

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau



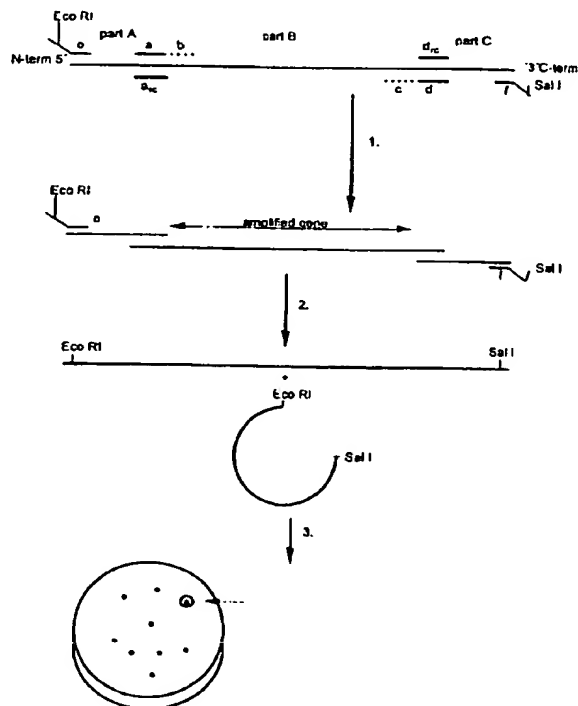
INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : C12N 15/10, 15/62, 15/56, 9/24, 9/42, C12Q 1/68		A2	(11) International Publication Number: WO 97/43409
			(43) International Publication Date: 20 November 1997 (20.11.97)
(21) International Application Number: PCT/DK97/00216		(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TR, TT, UA, UG, US, UZ, VN, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).	
(22) International Filing Date: 12 May 1997 (12.05.97)		<p>Published Without international search report and to be republished upon receipt of that report.</p>	
(30) Priority Data: 0562/96 10 May 1996 (10.05.96) DK			
(71) Applicant (for all designated States except US): NOVO NORDISK A/S [DK/DK]; Novo Allé, DK-2880 Bagsværd (DK).			
(72) Inventors; and (75) Inventors/Applicants (for US only): DALBØGE, Henrik [DK/DK]; Novo Nordisk A/S, Novo Allé, DK-2880 Bagsværd (DK). DIDERICHSEN, Børge [DK/DK]; Novo Nordisk A/S, Novo Allé, DK-2880 Bagsværd (DK). SANDAL, Thomas [DK/DK]; Novo Nordisk A/S, Novo Allé, DK-2880 Bagsværd (DK). KAUPPINEN, Sakari [FI/DK]; Novo Nordisk A/S, Novo Allé, DK-2880 Bagsværd (DK).			
(74) Common Representative: NOVO NORDISK A/S; Novo Allé, DK-2880 Bagsværd (DK).			

(54) Title: METHOD OF PROVIDING NOVEL DNA SEQUENCES

(57) Abstract

The present invention relates to a method of providing novel DNA sequences encoding a polypeptide with an activity of interest, comprising the following steps: i) PCR amplification of said DNA with PCR primers with homology to (a) known gene(s) encoding a polypeptide with an activity of interest, ii) linking the obtained PCR product to a 5' structural gene sequence and a 3' structural gene sequence, iii) expressing said resulting hybrid DNA sequence, iv) screening for hybrid DNA sequences encoding a polypeptide with said activity of interest or related activity, v) isolating the hybrid DNA sequence identified in step iv). Further, the invention also relates novel DNA sequences provided according to the method of the invention and polypeptides with an activity of interest encoded by said novel DNA sequences of the invention.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

Title: Method of providing novel DNA sequences

FIELD OF THE INVENTION

The present invention relates to a method of providing novel DNA sequences encoding a polypeptide with an activity of interest, novel DNA sequences provided according to the method of the invention, polypeptides with an activity of interest encoded by novel DNA sequences of the invention.

BACKGROUND OF THE INVENTION

10 The advent of recombinant DNA techniques has made it possible to select single protein components with interesting properties and produce them on a large scale. This represents an improvement over the previously employed production process using microorganisms isolated from nature and producing a mixture of proteins
15 which would either be used as such or separated after the production step.

Since the traditional methods were rather time-consuming, more rapid and less cumbersome methods were developed.

A such technique is described in WO 93/11249 (Novo Nordisk
20 A/S).

The method described in WO 93/11249 comprises the steps of:

- a) cloning, in suitable vectors, a DNA library from an organism suspected of producing one or more proteins of interest;
- b) transforming suitable yeast host cells with said vectors;
- 25 c) culturing the host cells under suitable conditions to express any protein of interest encoding by a clone in the DNA library; and
- d) screening for positive clones by determining any activity of a protein expressed in step c).

30 According to this method it is necessary to prepare a DNA library, comprising complete genes encoding polypeptides with activities of interest. Such a library has traditionally been made on mRNA isolated from micro-organisms which has been cultivated and isolated.

35 As it is only possible with known methods to cultivate about 2% of the microorganisms known today (i.e. cultivable microorganisms), genes encoding polypeptides from a huge number of

microorganisms (i.e. un-cultivable microorganisms) are generally difficult to identify and clone on the basis of screening technologies used today, such as the above mentioned.

5 SUMMARY OF THE INVENTION

It is the object of the present invention to provide a method for providing a novel DNA sequence encoding a polypeptide with an activity of interest from micro-organisms without having to cultivate and isolate said micro-organisms.

10 In the first aspect the invention relates to a method of providing novel DNA sequences encoding a polypeptide with an activity of interest, comprising the following steps:

- i) PCR amplification of said DNA with PCR primers with homology to (a) known gene(s) encoding a polypeptide with an activity of
15 interest,
- ii) linking the obtained PCR product to a 5' structural gene sequence and a 3' structural gene sequence,
- iii) expressing said resulting hybrid DNA sequence,
- iv) screening for hybrid DNA sequences encoding a polypeptide
20 with said activity of interest or related activity,
- v) isolating the hybrid DNA sequence identified in step iv)

Further, the invention also relates novel DNA sequences provided according to the method of the invention and polypeptides with an activity of interest encoded by said novel
25 DNA sequences of the invention.

BRIEF DESCRIPTION OF THE DRAWING

Figure 1 shows the cloning strategy of novel hybrid enzyme sequences.

- 30 a is an exact N-terminal consensus primer
a_{rc} is the reverse and complement primer to a
b is a degenerated homologous N-terminal primer
c is a degenerated homologous C-terminal primer
d is an exact C-terminal consensus primer
35 d_{rc} is a reverse and complement of d
f is an exact reverse and complement C-terminal primer extended with a sequence which includes a SalI restriction recognition site.

e is an exact N-terminal primer extended with a sequence which includes an EcoRI restriction recognition site.

1. (in figure 1)

PCR with primers ab and cd to amplify unknown core genes with an activity of interest.

PCR with primers e and a_{rc} to obtain the N-terminal part of the known gene.

PCR with primers d_{rc} and f to obtain the C-terminal part of the known gene.

2. (in figure 1)

SOE-PCR with primers e and f to link the unknown core gene sequence with the known N- and C-terminal gene sequences and introduction of EcoRI and SalI restriction recognition sites.

3. Restriction enzyme digestion followed by ligation of the novel sequence into an expression vector and transformation into a host cell. Screening of clones expressing the produced gene product with the activity of interest.

Figure 2 shows a part of an alignment of prokaryote xylanases belonging to glycosyl hydrolases family 11.

Figure 3 shows an alignment of the translated DNA sequences of Pulpzyme® (SEQ ID NO 2) and the novel gene sequence found in soil, respectively.

Figure 4 shows a schematically a novel hybrid gene provided according to the invention. Part A and Part C are the known sequences linked to the unknown Part B.

Using Pulpzyme® (SEQ ID NO 1) as the starting sequence:

"1" indicated the first nucleotide of the novel hybrid gene provided according to the invention, "433" and "631" the start and end of the part constituted by the unknown gene sequence and "741" the last nucleotide of the novel hybrid gene sequence.

DEFINITIONS

Prior to discussing this invention in further detail, the following terms will first be defined.

"Homology of DNA sequences or polynucleotides" In the present context the degree of DNA sequence homology is determined as the degree of identity between two sequences indicating a derivation of the first sequence from the second. The homology may suitably be determined by means of computer programs known in the art, such as GAP provided in the GCG program package (Program Manual for the Wisconsin Package, Version 8, August 1994, Genetics Computer Group, 575 Science Drive, Madison, Wisconsin, USA 53711)(Needleman, S.B. and Wunsch, C.D., (1970), Journal of Molecular Biology, 48, 443-453).

"Homologous": The term "homologous" means that one single-stranded nucleic acid sequence may hybridize to a complementary single-stranded nucleic acid sequence. The degree of hybridization may depend on a number of factors including the amount of identity between the sequences and the hybridization conditions such as temperature and salt concentration as discussed later (vide infra).

Using the computer program GAP (vide supra) with the following settings for DNA sequence comparison: GAP creation penalty of 5.0 and GAP extension penalty of 0.3, it is in the present context believed that two DNA sequences will be able to hybridize (using low stringency hybridization conditions as defined below) if they mutually exhibit a degree of identity preferably of at least 70%, more preferably at least 80%, and even more preferably at least 85%.

"heterologous": If two or more DNA sequences mutually exhibit a degree of identity which is less than above specified, they are in the present context said to be "heterologous".

"Hybridization:" Suitable experimental conditions for determining if two or more DNA sequences of interest do hybridize or not is herein defined as hybridization at low stringency as described in detail below.

A suitable experimental low stringency hybridization protocol between two DNA sequences of interest involves pre-soaking of a filter containing the DNA fragments to hybridize

in 5 x SSC (Sodium chloride/Sodium citrate, Sambrook et al. 1989) for 10 min, and prehybridization of the filter in a solution of 5 x SSC, 5 x Denhardt's solution (Sambrook et al. 1989), 0.5 % SDS and 100 µg/ml of denatured sonicated salmon sperm DNA (Sambrook et al. 1989), followed by hybridization in the same solution containing a concentration of 10ng/ml of a random-primed (Feinberg, A. P. and Vogelstein, B. (1983) *Anal. Biochem.* 132:6-13), ³²P-dCTP-labeled (specific activity > 1 x 10⁹ cpm/µg) probe (DNA sequence) for 12 hours at ca. 45°C. The filter is then washed twice for 30 minutes in 2 x SSC, 0.5 % SDS at least 50°C, more preferably at least 55°C, and even more preferably at least 60°C (high stringency).

Molecules to which the oligonucleotide probe hybridizes under these conditions are detected using a x-ray film.

"Alignment": The term "alignment" used herein in connection with a alignment of a number of DNA and/or amino acid sequences means that the sequences of interest is aligned in order to identify mutual/common sequences of homology/identity between the sequences of interest. This procedure is used to identify common "conserved regions" (vide infra), between sequences of interest. An alignment may suitably be determined by means of computer programs known in the art, such as ClusterW or PILEUP provided in the GCG program package (Program Manual for the Wisconsin Package, Version 8, August 1994, Genetics Computer Group, 575 Science Drive, Madison, Wisconsin, USA 53711)(Needleman, S.B. and Wunsch, C.D., (1970), *Journal of Molecular Biology*, 48, 443-453).

"Conserved regions:" The term "conserved region" used herein in connection with a "conserved region" between DNA and/or amino acid sequences of interest means a mutual common sequence region of the sequences of interest, wherein there is a relatively high degree of sequence identity between the sequences of interest. In the present context a conserved region is preferably at least 10 base pairs (bp)/ 3 amino acids(a.a), more preferably at least 20 bp/ 7 a.a., and even more preferably at least 30 bp/ 10 a.a..

Using the computer program GAP (Program Manual for the Wisconsin Package, Version 8, August 1994, Genetics Computer

Group, 575 Science Drive, Madison, Wisconsin, USA 53711) (Needleman, S.B. and Wunsch, C.D., (1970), Journal of Molecular Biology, 48, 443-453) (vide supra) with the following settings for DNA sequence comparison: GAP creation penalty of 5.0 and GAP extension penalty of 0.3, the degree of DNA sequence identity within the conserved region is preferably of at least 80%, more preferably at least 85%, more preferably at least 90%, and even more preferably at least 95%.

"Sequence overlap extension PCR reaction (SOE-PCR)": The term "SOE-PCR" is a standard PCR reaction protocol known in the art, and is in the present context defined and performed according to standard protocols defined in the art ("PCR A practical approach" IRL Press, (1991)).

"primer": The term "primer" used herein especially in connection with a PCR reaction is an oligonucleotide (especially a "PCR-primer") defined and constructed according to general standard specification known in the art ("PCR A practical approach" IRL Press, (1991)).

"A primer directed to a sequence:" The term "a primer directed to a sequence" means that the primer (preferably to be used in a PCR reaction) is constructed so it exhibits at least 80% degree of sequence identity to the sequence part of interest, more preferably at least 90% degree of sequence identity to the sequence part of interest, which said primer consequently is "directed to". The primer is designed in order to specifically anneal at the region at a given temperature it is directed towards. Especially identity at the 3' end of the primer is essential for the function of the polymerase, i.e. the ability of a polymerase to extend the annealed primer.

"Polypeptide" Polymers of amino acids sometimes referred to as protein. The sequence of amino acids determines the folded conformation that the polypeptide assumes, and this in turn determines biological properties such as activity. Some polypeptides consist of a single polypeptide chain (monomeric), whilst other comprise several associated polypeptides (multimeric). All enzymes and antibodies are polypeptides.

"Enzyme" A protein capable of catalysing chemical reactions. Specific types of enzymes are a) hydrolases

including amylases, cellulases and other carbohydrases, proteases, and lipases, b) oxidoreductases, c) Ligases, d) Lyases, e) Isomerases, f) Transferases, etc. Of specific interest in relation to the present invention are enzymes used
5 in detergents, such as proteases, lipases, cellulases, amylases, etc.

"known sequence" is the term used for the DNA sequences of which the full length sequence has been sequenced or at least the sequence of one conserved regions is known.

10 "unknown sequence" is the term used for the DNA sequences amplified directly from uncultivated micro-organisms comprised in e.g. a soil sample used as the starting materia. "Full length DNA sequence" means a structural gene sequence encoding a complete polypeptide with an activity of interest.

15 "un-cultivated" means that the micro-organism comprising the unknown DNA sequence need not be isolated (i.e. to provide a population comprising only identical micro-organisms) before amplification (e.g. by PCR).

The term "an activity of interest" means any activity for
20 which screening methods is known.

The term "un-cultivable micro-organisms" defined micro-organisms which can not be cultivated according to methods know in the art.

The term "DNA" should be interpreted as also covering other
25 polynucleotide sequences including RNA.

The term "linking" sequences means effecting a covalent binding of DNA sequences.

The term "hybrid sequences" means sequences of different origin merged together into one sequence.

30 The term "structural gene sequence" means a DNA sequence coding for a polypeptide with an activity.

The term "natural occurring DNA" means DNA, which has not been subjected to biological or biochemical mutagenesis. By biological mutagenesis is meant "in vivo" mutagenesis, i.e.
35 propagation under controlled conditions in a living organism, such as a "mutator" strain, in order to create genetic diversity. By biochemical mutagenesis is meant "in vitro" mutagenesis, such as error-prone PCR, oligonucleotide directed

site-specific or random mutagenesis etc.

DETAILED DESCRIPTION OF THE INVENTION

It is the object of the present invention to provide a method
5 for providing novel DNA sequences encoding polypeptides with an
activity of interest from micro-organisms without having to
cultivate said micro-organisms.

The inventors of the present invention have found that PCR-
screening using primers designed on the basis of known
10 homologous region, such as conserved regions, can be used for
providing novel DNA sequences. Despite the fact that known
homologous regions, such as conserved regions, are used for
primer designing a vast number of unknown DNA sequences have been
provided. This will be described in the following and illustrated
15 in the Examples.

The DNA sequences provided are full length hybrid structural
gene sequences encoding complete polypeptides with an activity of
interest made up of one unknown sequence and one or two known
sequences.

20 According to the invention it is essential to identify at
least two homologous regions, such as conserved regions, in known
gene sequences with the activity of interest. One or two selected
known structural gene sequence(s) is(are) used as templates (i.e.
as starting sequence(s)) for finding and constructing novel DNA
25 structural gene sequences with an activity of interest.

Said homologous regions, such as conserved regions, can be
identified by alignment of polypeptides with the activity of
interest and may e.g. be made by the computer program ClustalW
or other similar programs available on the market.

30

One known structural gene as the starting sequence

In the case of using one known structural gene sequence as the
starting sequence it will typically be comprised in a plasmid or
vector or the like. A part of the sequence between the two
35 identified homologous regions, such as conserved regions, are
deleted to avoid contamination by the wild-type structural gene.

The known DNA sequence, with the homologous regions, such as
conserved regions, placed at the ends, are linked to an unknown

DNA sequence amplified directly or indirectly from a sample comprising micro-organisms.

The identified homologous regions, such as conserved regions, must have a suitable distance from each other, such as 10 or more
5 base pairs in between. It is preferred to use homologous regions, such as conserved regions, placed in each end of the known structural full length gene.

However, if knowledge about a specific function (e.g. active site) of a domain (i.e. part of the structural gene sequence) is
10 available it may be advantageous to use conserved regions placed in proximity of and on each side said domain as basis for the PCR amplification to provide novel DNA sequences according to the invention which will be described below in details.

15 Two known genes as starting sequences

In the case of using two known structural genes as the starting sequences at least one homologous region, such as conserved region, should be identified in each of the two sequences within the polypeptide coding region.

20 In both case (i.e. one or two known genes as starting sequences) the homologous regions, such as conserve regions, should preferably be situated at each end of the structural gene(s) (i.e. the sequences encoding the N-terminal end (i.e. named Part A on figure 4) and the C-terminal end, respectively
25 (i.e. named Part C on figure 4) of the known part of the hybrid polypeptide

In the first aspect the invention relates to a method for providing novel DNA sequences encoding polypeptides with an activity of interest comprises the following steps:

- 30 i) PCR amplification of said DNA with PCR primers with homology to (a) known gene(s) encoding a polypeptide with an activity of interest,
ii) linking the obtained PCR product to a 5' structural gene sequence and a 3' structural gene sequence,
35 iii) expressing said resulting hybrid DNA sequence,
iv) screening for hybrid DNA sequences encoding a polypeptide with said activity of interest or related activity,

v) isolating the hybrid DNA sequence identified in step iv)

In step i) the part between the corresponding homologous regions, such as conserved regions, of the unknown structural gene are amplified.

5 In an embodiment the PCR amplification in step i) is performed using naturally occurring DNA or RNA as template.

In another embodiment the micro-organism has not been subjected to "in vitro" selection.

The PCR amplification may be performed on a sample containing
10 DNA or RNA from un-isolated micro-organisms. According to the invention no prior knowledge about the unknown sequence is required.

In an embodiment of the invention said 5' and 3' structural gene sequences originate from two different known structural gene
15 sequences encoding polypeptides having the same activity or related activity.

The 5' structural gene sequence and the 3' structural gene sequence may also originate from the same known structural gene encoding a polypeptide with the activity of interest or from two
20 different known structural gene sequences encoding polypeptides having different activities. In the latter case it is preferred that at least one of the starting sequences originates from a known structural gene sequence encoding a polypeptide with the activity of interest.

25 In a preferred embodiment of the method of the invention the known structural gene is situated in a plasmid or a vector. In said case the method comprises the following steps:

- i) PCR amplification of DNA from micro-organisms with PCR primers being homologous to conserved regions of
30 a known gene encoding a polypeptide with an activity of interest,
- ii) cloning the obtained PCR product into a gene encoding a polypeptide having said activity of interest, where said gene is not identical to the gene from which the
35 PCR product is obtained, which gene is situated in an expression vector,
- iii) transforming said expression vector into a suitable host cell,

- iiia) culturing said host cell under suitable conditions,
- iv) screening for clones comprising a DNA sequence originated from the PCR amplification in step i) encoding a polypeptide with said activity of interest or a related activity,
- 5 v) isolating the DNA sequence identified in step iv).

According to this embodiment one known structural gene sequence is used as the starting sequence. It is to be understood that the PCR product obtained in step i) is cloned into a known
10 gene where a part of the DNA sequence, between the conserved regions, is deleted (i.e. cut out) or in an other way substituted with the PCR product. The deleted part of the known gene comprised in the vector may have any suitable size, typically between 10 and 5000 bp, such as from between 10 to 3000 bp.

15 A general problem is that, when amplifying DNA sequences encoding polypeptides with an activity by PCR, the obtained PCR product (i.e. being a part of an unknown gene) does not normally encode a polypeptide with the desired activity of interest.

Therefore, according to the invention the complete full length
20 structural gene, encoding a functional polypeptide, is provided by cloning (i.e. by substituting) the PCR product of the unknown structural gene into the known gene situated on the expression vector.

It should be emphasised that the DNA mentioned in step i), to
25 be PCR amplified, need not to comprise a complete gene encoding a functional polypeptide. This is advantageous as only a smaller region of the DNA of the micro-organism(s) in question need to be amplified.

The novel DNA sequences obtained according to the invention
30 consist of the PCR product merged or linked into the known gene, having a number of nucleotides between the conserved regions deleted. The PCR product is inserted into the known gene between the two ends of the cut open vector by overlapping homologous regions of about 10 to 200 bp at each end of the vector.

35 The resulting novel hybrid DNA sequences constitute complete full length genes comprising the PCR product and encodes a polypeptide with the activity of interest.

It is to be understood that it is not absolutely necessary to delete a part of the known gene sequence. However, if a part of the known gene sequence is not deleted re-ligation results in that the wild-type activity of the known gene is regained and
5 thus give a high number of wild-type background clones, which would make the screening procedure more time consuming and cumbersome.

The PCR amplification in step i) can be performed on both cultivable and uncultivable micro-organisms by directly or
10 indirectly amplification of DNA from the genomic material of the micro-organisms in the environment (i.e. directly or indirectly from the sample taken).

The micro-organisms

15 The micro-organisms from which the unknown DNA sequences are derived may be micro-organisms which cannot today be cultivated. This is possible as the DNA sequences can be amplified by PCR without the need first to cultivate and isolate the micro-organisms comprising the unknown DNA sequence(s).

20 It is however to be understood that the method of the invention can also be used for providing novel DNA sequences derived from micro-organisms which can be cultivated.

Therefore the method of the invention can be performed on both cultivable and un-cultivable organisms as the micro-organisms in
25 question do not, according to the method of the invention, need to be cultivated and isolated from, e.g. the soil sample, comprising micro-organisms.

Starting material

30 The starting material, i.e. the sample comprising micro-organisms with the target unknown DNA sequences, may for instance be an environmental samples of plant or soil material, animal or insect dung, insect gut, animal stomach, a marine sample of sea or lake water, sewage, waste water, etc., comprising one or, as
35 in most case, a vast number of different cultivable and/or un-cultivable micro-organisms.

If the genomic material of the micro-organisms are readily accessible the PCR amplification may be performed directly on the

sample. In other cases a pre-purification and isolation procedure of the genomic material is needed.

Smalla et al. (1993), J. Appl. Bacteriol 74, p. 78-85; Smalla et al. (1993), FEMS Microbiol Ecol 13, p. 47-58, describes how to
5 extract DNA directly from micro-organisms in the environment (i.e. the sample).

Borneman et al. (1996), Applied and Environmental Microbiology, 1935-1943, describes a method for extracting DNA from soils.

10 A commercially available kit for isolating DNA from environmental samples, such as e.g. soils, can be purchased from BIO 101 under the tradename FastDNA® SPIN Kit.

Seamless™ Cloning kit (catalogue no. Stratagene 214400) is a commercial kit suitable for cloning of any DNA fragment into any
15 desired location e.g. a vector, without the limitation of naturally occurring restriction sites.

PCR amplification of DNA and/or RNA of micro-organisms in the environment is described by Erlich, (1989), PCR Technology. Principles and Applications for DNA Amplification, New
20 York/London, Stockton Press; Pillai, et al., (1991), Appl. Environ. Microbiol, 58, p. 2712-2722)

Other methods for PCR amplifying microbial DNA directly from a sample is described in Molecular Microbial Ecology Manual, (1995), Edited by Akkermans et al.. A suitable method for
25 microbial DNA from soil samples is described by Jan Dirk van Elsas et al., (1995), Molecular Microbial Ecology Manual 2.7.2, p. 1-10.

Stein et al., (1996), J. Bacteriol., Vol. 178, No. 2, p. 591-599, describes a method for isolating DNA from un-cultivated
30 prokaryotic micro-organisms and cloning DNA fragments therefrom.

The PCR primers being homologous to conserved regions of the known gene encoding a polypeptide with an activity of interest are synthesized according to standard methods known in the art
35 (see for instance EP 684 313 from Hoffmann-La Roche AG) on the basis of knowledge to conserved regions in the polypeptide with the activity of interest.

Said PCR primers may be identical to at least a part of the conserved regions of the known gene. However, said primers may advantageously be synthesized to differ in one or more positions.

Further, a number of different PCR primers homologous to the
5 conserved regions may be used at the same time in step i) of the method of the invention.

The cultivable or uncultivable micro-organisms may be both prokaryotic organisms such as bacteria, or eukaryotic organisms including algae, fungi and protozoa.

10 Examples of un-cultivable organisms include, without being limited thereto, extremophiles and planktonic marine organisms etc.

The group of cultivable organisms include bacteria, fungal organisms, such as filamentous fungi or yeasts.

15 In the case of using DNA from cultivable organisms the PCR amplification in step i) may be performed on one or more polynucleotides comprised in a vector, plasmid or the like, such as on a cDNA library.

Specific examples of "an activity of interest" include enzymatic
20 tic activity and anti-microbial activity.

In a preferred embodiment of the invention the activity of interest is an enzymatic activity, such as an activity selected from the group comprising of phosphatases oxidoreductases (E.C. 1), transferases (E.C. 2); hydrolases (E.C. 3), such as esterases
25 (E.C. 3.1), in particular lipases and phytase; such as glucosidases (E.C. 3.2), in particular xylanase, cellulases, hemicellulases, and amylase, such as peptidases (E.C. 3.4), in particular proteases; lyases (E.C. 4); isomerases (E.C. 5); ligases (E.C. 6).

30 The host cell used in step iii) may be any suitable cell which can express the gene encoding the polypeptide with the activity of interest. The host cells may for instance be a yeast, such as a strain of *Saccharomyces*, in particular *Saccharomyces cerevisiae*, or a bacteria, such as a strain of *Bacillus*, in
35 particular of *Bacillus subtilis*, or a strain *Escherichia coli*.

Clones found to comprise a DNA sequence originated from the PCR amplification in step i) may be screened for any activity of interest. Examples of such activities include enzymatic activity,

anti-microbial activity or biological activities.

The polypeptide with the activity of interest may then be tested for a desired performance under specific conditions and/or in combination with e.g. chemical compounds or agent. In the case
5 where the polypeptide is an enzyme e.g. the wash performance, textile dyeing, hair dyeing or bleaching properties, effect in feed or food may be assayed to identify polypeptides with a desired property.

10 Identification of conserved regions of prokaryote xylanases

Figure 2 shows an alignment of prokaryote xylanases from the family 11 of glycosyl hydrolases (B. Henrissat, Biochem J, 280:309-316 (1991)). There are several region where the amino acids are identical or almost identical, i.e. conserved
15 regions.

Examples of homologous regions or conserved regions in prokaryotic xylanases from family 11 of glycosyl hydrolases (B. Henrissat, (1991), Biochem J 280:309-316) are the sequence "DGTTYDIY" (SEQ ID NO 3) position 145-152, "EGYQSSC" (SEQ ID
20 NO. 4) position 200-206 in the upper polypeptide shown in figure 2.

Based on e.g. said regions degenerated PCR primers can be designed. These degenerated PCR primers can amplify unknown DNA sequences coding for polypeptides (i.e. referred to as PCR
25 products below) which are homologous to the known polypeptide(s) in question (i.e. SEQ ID NO 2) flanked by the conserved regions.

The PCR products obtained can be cloned into a plasmid and sequenced to check if they contain conserved regions and are
30 homologous to the known structural gene sequence(s).

A homologous PCR product is however not a guarantee that the sequence code for a part of a polypeptide having the desired activity of interest.

Therefore, according to the method of the invention one or
35 more steps selecting DNA sequences encoding polypeptides having the activity of interest follow the construction of the novel hybrid DNA sequences.

The unknown DNA sequences

When method of the invention is performed on DNA from samples of uncultivated organisms it is advantageous to screen
5 for gene products with the activity of interest.

A suitable method for doing this is to link the PCR products with a 5' sequence upstream the first conserved region DNA sequence and the 3' sequence downstream the second consensus, respectively, from the known gene sequence.

10 The product of the unknown gene sequence linked to an N-terminal and C-terminal part of a known gene product is then screened for the activity of interest.

The N-terminal and C-terminal parts can originate from the same gene product but it is not a prerequisite for activity.
15 The N-terminal and C-terminal parts may also originate from different gene products as long as they originate from the same polypeptide family e.g. the same glycosyl hydrolases.

A method to link the unknown gene sequence with the known sequences is to clone the PCR product into a known gene,
20 encoding a polypeptide having the activity of interest, which have had the sequences between the conserved regions removed.

Another method is merging the PCR product, the N-terminal part and the C-terminal part by SOE-PCR (splicing by overlap extension PCR) e.g. as shown in figure 1 and described in
25 detail in Example 1. Other methods known in the art may also be used.

In a second aspect the invention relates to a novel DNA sequence provided by the method of the invention and the polypeptide encoded by said novel DNA sequence.

30

MATERIALS AND METHODS

Pulpzyme® is a xylanase derived from *Bacillus sp.* AC13, NCIMB No. 40482. and is described in WO 94/01532 from Novo Nordisk A/S
AZCL Birch xylan (MegaZyme, Australia).

35

Plasmids:

The *Aspergillus* expression vector pHD414 is a derivative of the plasmid p775 (described in EP 238 023). The construction of

pHD414 is further described in WO 93/11249.

The 43 kD EG V endoglucanase cDNA from *H. insolens* (disclosed in WO 91/17243) is cloned into pHD414 in such a way that the endoglucanase gene is transcribed from the TAKA-pro-
5 moter. The resulting plasmid is named pCaHj418.

Kits

QIAquick PCR Purification Kit Protocol
Tag deoxy terminal cycle sequencing kit (Perkin Elmer, USA)
10 AmpliTaq Gold polymerase (Perkin-Elmer, USA)

Micro-organisms

Bacteria

electromax DH10B *E. coli* cells (GIBCO BRL)
15

Fungal micro-organisms:

Cylindrocarpon sp.: Isolated from marine sample, the Bahamas

Classification: Ascomycota, Pyrenomycetes, Hypocreales
20 unclassified

Fusarium anguioides Sherbakoff IFO 4467

Classification: Ascomycota, Pyrenomycetes, Hypocreales, Hypocreaceae

Gliocladium catenulatum Gillman & Abbott CBS 227.48

25 Classification: Ascomycota, Pyrenomycetes, Hypocreales, Hypocreaceae

Humicola nigrescens Omvik CBS 819.73

Classification: Ascomycota, Pyrenomycetes, Sordariales, (fam. unclassified)

30 *Trichothecium roseum* IFO 5372

Plates

LB-ampicillin plates: 10 g Bacto-tryptone, 5 g Bacto yeast extract, 10 g NaCl, in 1 litre water, 2% agar 0.1% AZCL Birch
35 xylan, 50 microg/ml ampicillin.

Equipment

Applied Biosystems 373A automated sequencer

PCR Amplification

All Polymerase Chain Reactions is carried out under standard conditions as recommended by Perkin-Elmer using AmpliTag Gold polymerase.

Isolation of Environmental DNA

DNA is isolated from an environmental sample using FastDNA® SPIN Kit for Soil according to the manufacture's instructions.

Methods used in Example 3

Strains and growth conditions

The fungal strains listed above, were streaked on PDA plates containing 0.5 % Avicel, and examined under a microscope to avoid obvious mistakes and contaminations. The strains were cultivated in shake flasks (125 rpm and 26 °C) containing 30ml PD medium (to initiate the growth) and 150ml of BA growth medium for cellulase induction.

The production of cellulases in culture supernatants (typically after 3, 5, 7 and 9 days of growth) was assayed using 0.1 % AZCl-HE-cellulose in a plate assay at pH 3, pH 7 and pH 10. The mycelia were harvested and stored at - 80°C.

Preparation of RNase-free glassware, tips and solutions

All glassware used in RNA isolations were baked at + 250°C for at least 12 hours. Eppendorf tubes, pipet tips and plastic columns were treated in 0.1 % diethylpyrocarbonate (DEPC) in EtOH for 12 hours, and autoclaved. All buffers and water (except Tris-containing buffers) were treated with 0.1 % DEPC for 12 hours at 37°C, and autoclaved.

Extraction of total RNA

The total RNA was prepared by extraction with guanidinium thiocyanate followed by ultracentrifugation through a 5.7 M CsCl cushion [Chirgwin, (1979) Biochemistry 18, 5294-5299] using the following modifications. The frozen mycelia was ground in liquid N2 to fine powder with a mortar and a pestle,

followed by grinding in a precooled coffee mill, and immediately suspended in 5 vols of RNA extraction buffer (4 M GuSCN, 0.5 % Na-laurylsarcosine, 25 mM Na-citrate, pH 7.0, 0.1 M β -mercaptoethanol). The mixture was stirred for 30 min. at RT° and centrifuged (20 min., 10 000 rpm, Beckman) to pellet the cell debris. The supernatant was collected, carefully layered onto a 5.7 M CsCl cushion (5.7 M CsCl, 0.1 M EDTA, pH 7.5, 0.1 % DEPC; autoclaved prior to use) using 26.5 ml supernatant per 12.0 ml CsCl cushion, and centrifuged to obtain the total RNA (Beckman, SW 28 rotor, 25 000 rpm, RT°, 24h). After centrifugation the supernatant was carefully removed and the bottom of the tube containing the RNA pellet was cut off and rinsed with 70 % ETOH. The total RNA pellet was transferred into an Eppendorf tube, suspended in 500 μ l TE, pH 7.6 (if difficult, heat occasionally for 5 min at 65 °C), phenol extracted and precipitated with ethanol for 12 h at -20°C (2.5 vols ETOH, 0.1 vol 3M NaAc, pH 5.2). The RNA was collected by centrifugation, washed in 70 % ETOH, and resuspended in a minimum volume of DEPC-DIW. The RNA concentration was determined by measuring OD 260/280.

20

Isolation of poly(A)+RNA

The poly(A)+ RNAs were isolated by oligo(dT)-cellulose affinity chromatography [Aviv, (1972), Proc. Natl. Acad. Sci. U.S.A. 69, 1408-1412]. Typically, 0.2 g of oligo(dT) cellulose (Boehringer Mannheim, Germany) was preswollen in 10 ml of 1 x column loading buffer (20 mM Tris-Cl, pH 7.6, 0.5 M NaCl, 1 mM EDTA, 0.1 % SDS), loaded onto a DEPC-treated, plugged plastic column (Poly Prep Chromatography Column, Bio Rad), and equilibrated with 20 ml 1 x loading buffer. The total RNA (1-2 mg) was heated at 65 °C for 8 min., quenched on ice for 5 min, and after addition of 1 vol 2 x column loading buffer to the RNA sample loaded onto the column. The eluate was collected and reloaded 2-3 times by heating the sample as above and quenching on ice prior to each loading. The oligo(dT) column was washed with 10 vols of 1 x loading buffer, then with 3 vols of medium salt buffer (20 mM Tris-Cl, pH 7.6, 0.1 M NaCl, 1 mM EDTA, 0.1 % SDS), followed by elution of the poly(A)+ RNA with 3 vols of elution buffer (10 mM Tris-Cl, pH 7.6, 1 mM EDTA, 0.05% SDS)

preheated to + 65 °C, by collecting 500 µl fractions. The OD260 was read for each collected fraction, and the mRNA containing fractions were pooled and ethanol precipitated at -20°C for 12 h. The poly(A)+ RNA was collected by centrifugation, resuspended in DEPC-DIW and stored in 5-10 µg aliquots at -80 °C.

cDNA synthesis

First strand synthesis

Double-stranded cDNA was synthesized from 5 µg of poly(A)+ RNA by the RNase H method (Gubler et al. (1983) Gene 25, 263-269; Sambrook et al. (1989), Molecular Cloning: A Laboratory Manual, 2 Ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, New York) using the hair-pin modification. The poly(A)+RNA (5 µg in 5 µl of DEPC-treated water) was heated at 70°C for 8 min. in a pre-siliconized, RNase-free Eppendorph tube, quenched on ice, and combined in a final volume of 50 µl with reverse transcriptase buffer (50 mM Tris-Cl, pH 8.3, 75 mM KCl, 3 mM MgCl₂, 10 mM DTT, Bethesda Research Laboratories) containing 1 mM of dATP, dGTP and dTTP, and 0.5 mM of 5-methyl-dCTP (Pharmacia), 40 units of human placental ribonuclease inhibitor (RNasin, Promega), 1.45 µg of oligo(dT)18- Not I primer (Pharmacia) and 1000 units of SuperScript II RNase H- reverse transcriptase (Bethesda Research Laboratories). First-strand cDNA was synthesized by incubating the reaction mixture at 45 °C for 1 h. After synthesis, the mRNA:cDNA hybrid mixture was gel filtrated through a MicroSpin S-400 HR (Pharmacia) spin column according to the manufacturer's instructions.

Second strand synthesis

After the gel filtration, the hybrids were diluted in 250 µl of second strand buffer (20 mM Tris-Cl, pH 7.4, 90 mM KCl, 4.6 mM MgCl₂, 10 mM (NH₄)₂SO₄, 0.16 mM βNAD+) containing 200 µM of each dNTP, 60 units of *E. coli* DNA polymerase I (Pharmacia), 5.25 units of RNase H (Promega) and 15 units of *E. coli* DNA ligase (Boehringer Mannheim). Second strand cDNA synthesis was performed by incubating the reaction tube at 16°C for 2 h, and an additional 15 min at 25°C. The reaction was stopped by addition of EDTA to 20 mM final concentration followed by phenol

and chloroform extractions.

Mung bean nuclease treatment

The double-stranded (ds) cDNA was ethanol precipitated at -20°C for 12 hours by addition of 2 vols of 96% EtOH, 0.2 vol 10 M NH₄Ac, recovered by centrifugation, washed in 70% EtOH, dried (SpeedVac), and resuspended in 30 µl of Mung bean nuclease buffer (30 mM NaAc, pH 4.6, 300 mM NaCl, 1 mM ZnSO₄, 0.35 mM DTT, 2 % glycerol) containing 25 units of Mung bean nuclease (Pharmacia). The single-stranded hair-pin DNA was clipped by incubating the reaction at 30°C for 30 min, followed by addition of 70 µl 10 mM Tris-Cl, pH 7.5, 1 mM EDTA, phenol extraction, and ethanol precipitation with 2 vols of 96% EtOH and 0.1 vol 3M NaAc, pH 5.2 on ice for 30 min.

15 Blunt-ending with T4 DNA polymerase

The ds cDNAs were recovered by centrifugation (20 000 rpm, 30 min.), and blunt-ended with T4 DNA polymerase in 30 µl of T4 DNA polymerase buffer (20 mM Tris-acetate, pH 7.9, 10 mM MgAc, 50 mM KAc, 1 mM DTT) containing 0.5 mM each dNTP and 5 units of T4 DNA polymerase (New England Biolabs) by incubating the reaction mixture at +16°C for 1 hour. The reaction was stopped by addition of EDTA to 20 mM final concentration, followed by phenol and chloroform extractions and ethanol precipitation for 12 h at -20°C by adding 2 vols of 96% EtOH and 0.1 vol of 3M NaAc, pH 5.2.

Adaptor ligation, Not I digestion and size selection

After the fill-in reaction the cDNAs were recovered by centrifugation as above, washed in 70% EtOH, and the DNA pellet was dried in SpeedVac. The cDNA pellet was resuspended in 25 µl of ligation buffer (30 mM Tris-Cl, pH 7.8, 10 mM MgCl₂, 10 mM DTT, 0.5 mM ATP) containing 2.5 µg non-palindromic BstXI adaptors (1 µg/µl, Invitrogen) and 30 units of T4 ligase (Promega) by incubating the reaction mix at +16°C for 12 h. The reaction was stopped by heating at + 65°C for 20 min, and then on ice for 5 min. The adapted cDNA was digested with Not I restriction enzyme by addition of 20 µl autoclaved water, 5 µl of 10 x Not I restriction enzyme buffer (New England Biolabs) and 50 units

of Not I (New England Biolabs), followed by incubation for 2.5 hours at +37°C. The reaction was stopped by heating the sample at +65°C for 10 min. The cDNAs were size-fractionated by agarose gel electrophoresis on a 0.8% SeaPlaque GTG low melting temperature agarose gel (FMC) in 1 x TBE (in autoclaved water) to separate unligated adaptors and small cDNAs. The gel was run for 12 hours at 15 V, the cDNA was size-selected with a cut-off at 0.7 kb by cutting out the lower part of the agarose gel, and the cDNA was concentrated by running the gel backwards until it appeared as a compressed band on the gel. The cDNA (in agarose) was cut out from the gel, and the agarose was melted at 65°C in a 2 ml Biopure Eppendorph tube (Eppendorph). The sample was treated with agarase by adding 0.1 vol of 10 x agarase buffer (New England Biolabs) and 2 units per 100 µl molten agarose to the sample, followed by incubation at 45°C for 1.5 h. The cDNA sample was phenol and chloroform extracted, and precipitated by addition of 2 vols of 96 % EtOH and 0.1 vol of 3M NaAc, pH 5.2 at - 20°C for 12 h.

20 EXAMPLES

Example 1

Providing novel DNA sequences encoding polypeptide with xylanase activity

Novel sequences with xylanase activity were provided according to the method of the invention using the glycosyl hydrolase family 11 xylanase derived from *Bacillus sp.* (SEQ ID No 1) as the known structural gene sequence.

Identification of conserved regions by alignment

30 An amino acid sequence alignment of ten family 11 xylanases revealed at least 3 conserved sequences. Two of these conserved sequences are used to design appropriate PCR primers for amplification of unknown DNA sequences.

The first conserved sequence shown in SEQ ID No. 3 i.e. "DGGTYDIY" corresponding to position 433-456 in SEQ ID NO 1.

The second conserved sequence shown in SEQ 4, i.e. "EGYQSSG" corresponding to position 631-651 in SEQ ID NO 1.

PCR amplification of the known and unknown partial structural gene sequences

Initially the N-terminal end (i.e. Part A) and the C-terminal (i.e. Part C) of the known xylanase gene, in which the
5 unknown sequence (i.e. Part B) is to be inserted, were amplified by PCR (see figure 4)

Part A was PCR amplified using the two primers (i.e. primer e and primer arc) and as DNA template a plasmid carrying the known xylanase gene (i.e. SEQ ID NO 1).

10 Primer e (shown in SEQ ID NO 5 and figure 1) is an exact N-terminal primer extended with a sequence which included an EcoRI restriction recognition site.

Primer arc (shown in SEQ ID NO 6 and figure 1) is a reverse and complement sequence primer of position 411-432 in SEQ ID NO
15 1.

Part C was PCR amplified using the two primers (i.e. primer f and primer d_{rc}) mentioned below and as DNA template a plasmid carrying the known xylanase gene.

Primer f is an exact reverse and complement C-terminal primer extended with a sequence which having a SalI restriction recognition site is shown in SEQ ID No. 7.

Primer d_{rc} (SEQ ID NO 8) was designed on the basis of position 651-672 in SEQ ID No. 1.

Part B was PCR amplified using two primers (i.e. primer ab and primer cd) and as DNA template DNA purified from a soil
25 sample using the FastDNA® SPIN Kit.

Primer ab (SEQ ID NO 9) has the exact sequence of position 411-432 in SEQ ID 1 extended with degenerated xylanase consensus sequence covering position 433-452 in SEQ ID NO 1

30 Primer cd (SEQ ID NO: 10) has the exact reverse and complement sequence of position 672-651 in SEQ ID NO 1 extended with degenerated xylanase consensus sequence covering position 650-631 in SEQ ID NO 1.

The N-terminal part of the known xylanase gene (Part A) was
35 PCR amplified for 9 min. at 94°C followed by 30 cycles (45 second at 94°C, 45 seconds at 50°C and 1 min. at 72°C) and finally for 7 min. at 72°C. This gave a PCR product of approx. 450 bp.

The C-terminal part (Part C) of the known xylanase gene was PCR amplified for 9 min. at 94°C followed by 30 cycles (45 seconds at 94°C, 45 seconds at 50°C and 1 min. at 72°C) and finally for 7 min. at 72°C. This gave a PCR product of approx. 100 bp.

The unknown sequences (Part B) was PCR amplified for 9 min. at 94°C followed by 40 cycles (45 seconds at 94°C, 45 seconds at 50°C and 1 min. at 72°C) and finally for 7 min. at 72°C. This gave a PCR product of approx. 260 bp.

10 The PCR products mentioned above were carefully purified to avoid remains of template DNA which can produce false positive bands in the following SOE-PCR where the products are joined together to form hybrid sequences.

15 Construction of hybrid sequences

Hybrid sequences containing the N- and C-terminal parts of the known xylanase gene with core part of unknown genes was constructed by splicing by overlap extension PCR (SOE-PCR).

20 Equal molar amounts of Part A, Part B and Part C PCR products were mixed and PCR amplified under standard conditions except that the reaction was started without any primers.

The reaction started with 9 min. at 94°C followed by 4 cycles (45 seconds at 94°C, 45 seconds at 50°C, 1 min. at 72°C), then primers e and f (SEQ ID No. 5 and 7, respectively) 25 were added, followed by 25 cycles (45 seconds at 94°C, 45 seconds at 50°C, 1 min. at 72°C) and finally 7 min. at 72°C. This gave a SOE-PCR product of the expected size of approx. 770 bp.

30 Cloning of the hybrids

The SOE-PCR product was purified using the QIAquick PCR Purification Kit Protocol and digested overnight with EcoRI and SalI according to the manufacturers recommendation. The digested product was then ligated into an E. coli expression 35 vector overnight at 16°C (in this case a vector where the hybrid gene is under control of a temperature sensitive lambda repressor promoter).

The ligation mixture was transformed into electromax DH10B *E. coli* cells (GIBCO BRL) and plated on LB-ampicillin plates containing 0.1% AZCL Birch xylan. After induction of the promoter (by increasing the temperature to 42°C) xylanase positive colonies were identified as colonies surrounded by a blue halo.

Plasmid DNA was isolated from positive *E. coli* colonies using standard procedures and sequenced with the Taq deoxy terminal cycle sequencing kit (Perkin Elmer, USA) using an Applied Biosystems 373A automated sequencer according to the manufacturer's instructions.

The sequence of a positive clone is shown in SEQ ID NO 11 and the corresponding protein sequence is shown in SEQ ID NO 12.

An alignment of the known xylanase sequence (SEQ ID NO 2) and the novel DNA sequence provided according to the method of the invention can be seen in Figure 3. As can be seen the two protein sequences differ between the two identified conserved regions (i.e. SEQ ID NO 3 and SEQ ID NO 4, respectively).

Example 2

Efficiency of the method of the invention

Degenerated primers were designed on the basis of conserved regions identified by alignment of a number of family 5 cellulases and family 10 and 11 xylanases found on the Internet in ExPASy under Prosite (Dictionary of protein sites and patterns).

PCR amplification of a number of unknown structural gene sequences from soil and cow rumen samples were performed with various degenerated primers covering identified conserved region sequences to show how effective the method of the invention is.

The PCR products were cloned into the vector pCRtmII, provided with the original TA cloning kit from Invitrogen. Said vector provides the possibility to make blue-white screening, the white colonies were selected and the inserts were sequenced.

When editing the Sequence Listing below all sequences outside the two EcoRI sites in the polylinker were removed.

Therefore all sequences have a small additional part of the polylinker (i.e. from the EcoRI site to the TT overhang) in both ends of the sequences. These extensions are "GAATTCGGCT" and "AAGCCG".

5 1. PCR primers were designed on the basis of identified conserved regions #1 GWNLGN and #2 (E/D)HLIFE of cellulases from the glycosyl hydrolase family 5 aiming to provide novel sequences with cellulase activity.

SEQ ID NO 13 and 14 show the sequences obtained from a soil
10 sample. SEQ ID NO 15 and 16 show the sequences obtained from a cow rumen sample.

2. PCR primers were designed on the basis of identified conserved regions #1 GWNLGN and #3 RA(S/T)GGNN of cellulases from the glycosyl hydrolase family 5 aiming to provide novel
15 sequences with cellulase activity.

SEQ ID NO 17 to 19 show the sequences obtained from a cow rumen sample.

3. PCR primers were designed on the basis of identified conserved regions #2 (E/D)HLIFE and #3 RA(S/T)GGNN of cellula-
20 ses from the glycosyl hydrolase family 5 aiming to provide novel sequences with cellulase activity.

SEQ ID NO 20 to 22 show the sequences obtained from a cow rumen sample.

4. PCR primers were designed on the basis of identified
25 conserved regions #4 HTLVWH and #5 WDVVNE of xylanases from the glycosyl hydrolase family 10 aiming to provide novel sequences with xylanase activity.

SEQ ID NO 23 to 28 show the sequences obtained from a cow rumen sample.

30 5. PCR primers were designed on the basis of the identified conserved regions #4 HTLVWH and #6 (F/Y)(I/Y)NDYN of xylanases from the glycosyl hydrolase family 10 aiming to provide novel sequences with xylanase activity.

SEQ ID NO 29 to 33 show the sequences obtained from a cow rumen
35 sample.

6. PCR primers were designed on the basis of the identified conserved regions #5 WDVVNE and #6 (F/Y)(I/Y)NDYN of xylanases from the glycosyl hydrolase family 10 aiming to provide novel

sequences with xylanase activity.

SEQ ID NO 34 to 36 show the sequences obtained from a soil sample. SEQ ID NO 37 to 45 show the sequences obtained from a cow rumen sample

5 7. PCR primers were designed on the basis of the identified conserved regions #8 DGTTYDIY and #9 EGYQSSG of xylanases from the glycosyl hydrolase family 11 aiming to provide novel sequences with xylanase activity.

10 SEQ ID NO 46 to 49 show the sequences obtained from a soil sample. SEQ ID NO 50 to 54 show the sequences obtained from a cow rumen sample.

60 clones with inserts were sequenced and resulted in 43 different sequences all encoding either a part of a cellulase or a part of a xylanase. Only 2 of the 43 sequences were 15 similar to sequence found in the sequence databases Genbank.

SEQ ID NO 49 was found to be similar to Xylanase A from *Bacillus pumilus*. SEQ ID NO 42 was found to be similar to a xylanase from *Prevotella ruminicola*.

20 Example 3

Construction of novel hybrid DNA sequences encoding polypeptides with endoglucanase activity

Novel hybrid DNA sequences with endoglucanase activity were provided by first identifying two conserved regions common for 25 the following family 45 cellulases (see WO 96/29397): *Humicola insolens* EGV (disclosed in WO 91/17243), *Fusarium oxysporum* EGV (Sheppard et al., Gene (1994), Vol. 15, pp.163-167), *Thielavia terrestris*, *Myceliophthora thermophila*, and *Acremonium* sp (disclosed in WO 96/29397).

30 The amino acid sequence alignment revealed two conserved region.

The first conserved region "Thr Arg Tyr Trp Asp Cys Cys Lys Pro/Thr" shown in SEQ ID NO 57 corresponds to position 6 to 14 of SEQ ID NO 55 showing the *Humicola insolens* EG V 43 KDa 35 endoglucanase.

The second conserved region "Trp Arg Phe/Tyr Asp Trp Phe" shown in SEQ ID NO 58 corresponding to positions 169 to 198 of SEQ ID NO 55 showing the *Humicola insolens* EGV 43 KDa

endoglucanase.

Two degenerate, deoxyinosine-containing oligonucleotide primers (sense; primer s and antisense; primer as) were constructed) for PCR amplification of unknown gene sequences. The 5 deoxyinosines are depicted by an I in the primer sequences.

Primers s and primer as are shown in SEQ ID No. 59 and 60 respectively.

The *Humicola insolens* EG V structural gene sequence (SEQ ID NO 55) was used as the known DNA sequence. A number of fungal 10 DNA sequences mentioned below were used as the unknown sequences.

PCR cloning of the family 45 cellulase core region and the linker/CBD of *Humicola insolens* EG V.

15 Approximately 10 to 20 ng of double-stranded, cellulase-induced cDNA from *Humicola nigrescens*, *Cylindrocarpon* sp., *Fusarium anguioides*, *Gliocladium catenulatum*, and *Trichothecium roseum* prepared, as described above in the Material and Methods section were, PCR amplified in Expand buffer (Boehringer Mannheim, Germany) containing 200 μ M each dNTP and 200 pmol of each 20 degenerate Primer s (SEQ ID NO 59) and Primer as (SEQ ID NO 60) a DNA thermal cycler (Perkin-Elmer, Cetus, USA) and 2.6 units of Expand High Fidelity polymerase (Boehringer Mannheim, Germany). 30 cycles of PCR were performed using a cycle profile of 25 denaturation at 94°C for 1 min, annealing at 55°C for 2 min, and extension at 72°C for 3 min, followed by extension at 72°C for 5 min.

The PCR fragment coding for the linker/CBD of *H. insolens* EGV was generated in Expand buffer (Boehringer Mannheim, Germany) containing 200 μ M each dNTP using 100 ng of the pCaHj418 30 many) template, 200 pmol forward primer 1 (SEQ ID NO 61), 200 pmol reverse primer 1 (SEQ ID NO 62). 30 cycles of PCR were performed as above.

35 Construction of hybrid genes using splicing by overlap extension (SOE)

The PCR products were electrophoresed in 0.7 % agarose gels (SeaKem, FMC), the fragments of interest were excised from the

gel and recovered by Qiagen gel extraction kit (Qiagen, USA) according to the manufacturer's instructions. The recombinant hybrid genes were generated by combining the overlapping PCR fragments from above (ca. 50 ng of each template) in Expand 5 buffer (Boehringer Mannheim, Germany) containing 200 μ M each dNTP in the SOE reaction. Two cycles of PCR were performed using a cycle profile of denaturation at 94°C for 1 min, annealing at 50 C for 2 min, and extension at 72°C for 3 min, the reaction was stopped, 250 pmol of each end-primer: forward 10 primer 2 (SEQ ID NO 63) encoding the TAKA-amylase signal sequence from *A. oryzae*, reverse primer 2 (SEQ ID NO 64) was added to the reaction mixture, and an additional 30 cycles of PCR were performed using a cycle profile of denaturation at 94°C for 1 min, annealing at 55 °C for 2 min, and extension at 72°C 15 for 3 min.

Construction of the expression cassettes and heterologous expression in *Aspergillus oryzae*

The PCR-generated, recombinant fragments were electropho- 20 resed in 0.7 % agarose gels (SeaKem, FMC), the fragments were excised from the gel and recovered by Qiagen gel extraction kit (Qiagen, USA) according to the manufacturer's instructions. The DNA fragments were digested to completion with BamHI and XbaI, and ligated into BamHI/XbaI-cleaved pHD414 vector. Co-transfor- 25 mation of *A. oryzae* was carried out as described in Christensen et al. (1988), Bio/Technology 6, 1419-1422. The AmdS+ transformants were screened for cellulase activity using 0.1 % AZCl-HE-cellulose in a plate assay as described above. The cellulase-producing transformants were purified twice through conidial 30 spores, cultivated in 250 ml shake flasks, and the amount of secreted cellulase was estimated by SDS-PAGE, Western blot analysis and the activity assay as described earlier (Kauppinen et al. (1995), J. Biol. Chem. 270, 27172-27178;; Kofod et al. (1994), J. Biol. Chem. 269, 29182-29189; Christgau et. 35 al, (1994), Biochem. Mol. Biol. Int. 33, 917 - 925).

Nucleotide sequence analysis

The nucleotide sequences of the novel hybrid gene fusions were determined from both strands by the dideoxy chain-termination method (Sanger et al., (1977), Proc. Natl. Acad. Sci. U.S.A. 74, 5463-5467), using 500 ng template, the Taq
5 deoxy-terminal cycle sequencing kit (Perkin-Elmer, USA), fluorescent labeled terminators and 5 pmol of synthetic oligonucleotide primers. Analysis of the sequence data was performed according to Devereux et al., 1984 (Devereux et al., (1984), Nucleic Acids Res. 12, 387-395).

10 The provided novel hybrid DNS sequences and the deduced protein sequences are shown in SEQ ID NO 65 to 74.

SEQ ID NO 65 shows the hybrid gene construct comprising the family 45 cellulase core region from *Humicola nigrescens* and the linker/CBD of *Humicola insolens* EG V. SEQ. ID No 66 shows
15 the deduced amino acid sequence of the hybrid gene construct.

SEQ ID NO 67 shows the hybrid gene construct comprising the family 45 cellulase core region from *Cylindrocarpum* sp. and the linker/CBD of *Humicola insolens* EG V. SEQ ID NO 68 shows the deduced amino acid sequence of the hybrid gene construct.

20 SEQ ID NO 69 shows the hybrid gene construct comprising the family 45 cellulase core region from *Fusarium anguioides* and the linker/CBD of *Humicola insolens* EG V. SEQ ID NO 70 shows the deduced amino acid sequence of the hybrid gene construct.

SEQ ID NO 71 shows the hybrid gene construct comprising the
25 family 45 cellulase core region from *Gliocladium catenulatum* and the linker/CBD of *Humicola insolens* EG V. SEQ ID NO 72 shows the deduced amino acid sequence of the hybrid gene construct.

SEQ ID NO 73 shows the novel gene construct comprising the
30 family 45 cellulase core region from *Trichothecium roseum* and the linker/CBD of *Humicola insolens* EG V. SEQ ID NO 74 shows the deduced amino acid sequence of the hybrid gene construct.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT:

- 5 (A) NAME: Novo Nordisk A/S
 (B) STREET: Novo Alle
 (C) CITY: Bagsvaerd
 (E) COUNTRY: Denmark
 (F) POSTAL CODE (ZIP): DK-2880
 10 (G) TELEPHONE: +45 4444 8888
 (H) TELEFAX: +45 4449 3256
 (ii) TITLE OF INVENTION: Method for providing novel DNA sequences
 (iii) NUMBER OF SEQUENCES: 74
 (iv) COMPUTER READABLE FORM:
 15 (A) MEDIUM TYPE: Floppy disk
 (B) COMPUTER: IBM PC compatible
 (C) OPERATING SYSTEM: PC-DOS/MS-DOS
 (D) SOFTWARE: PatentIn Release #1.0, Version #1.30 (EPO)

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 747 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 25 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: DNA (genomic)
 (vi) ORIGINAL SOURCE:
 (B) STRAIN: Bacillus sp. AC13, NCIMB No. 40482
 (ix) FEATURE:
 30 (A) NAME/KEY: CDS
 (B) LOCATION: 1..747
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

35	ATG AGA CAA AAG AAA TTG ACG TTC ATT TTA GCC TTT TTA GTT TGT TTT	48
	Met Arg Gln Lys Lys Leu Thr Phe Ile Leu Ala Phe Leu Val Cys Phe	
	1 5 10 15	
40	GCA CTA ACC TTA CCT GCA GAA ATA ATT CAG GCA CAA ATC GTC ACC GAC	96
	Ala Leu Thr Leu Pro Ala Glu Ile Ile Gln Ala Gln Ile Val Thr Asp	
	20 25 30	
45	AAT TCC ATT GGC AAC CAC GAT GGC TAT GAT TAT GAA TTT TGG AAA GAT	144
	Asn Ser Ile Gly Asn His Asp Gly Tyr Asp Tyr Glu Phe Trp Lys Asp	
	35 40 45	
50	AGC GGT GGC TCT GGG ACA ATG ATT CTC AAT CAT GGC GGT ACG TTC AGT	192
	Ser Gly Gly Ser Gly Thr Met Ile Leu Asn His Gly Gly Thr Phe Ser	
	50 55 60	
55	GCC CAA TGG AAC AAT GTT AAC AAC ATA TTA TTC CGT AAA GGT AAA AAA	240
	Ala Gln Trp Asn Asn Val Asn Asn Ile Leu Phe Arg Lys Gly Lys Lys	
	65 70 75 80	
60	TTC AAT GAA ACA CAA ACA CAC CAA CAA GTT GGT AAC ATG TCC ATA AAC	288
	Phe Asn Glu Thr Gln Thr His Gln Gln Val Gly Asn Met Ser Ile Asn	
	85 90 95	
65	TAT GGC GCA AAC TTC CAG CCA AAC GGA AAT GCG TAT TTA TGC GTC TAT	336
	Tyr Gly Ala Asn Phe Gln Pro Asn Gly Asn Ala Tyr Leu Cys Val Tyr	
	100 105 110	
70	GGT TGG ACT GTT GAC CCT CTT GTC GAA TAT TAT ATT GTC GAT AGT TGG	384
	Gly Trp Thr Val Asp Pro Leu Val Glu Tyr Tyr Ile Val Asp Ser Trp	
	115 120 125	
75	GGC AAC TGG CGT CCA CCA GGG GCA ACG CCT AAG GGA ACC ATC ACT GTT	432
	Gly Asn Trp Arg Pro Pro Gly Ala Thr Pro Lys Gly Thr Ile Thr Val	
	130 135 140	

32

	GAT	GGA	GGA	ACA	TAT	GAT	ATC	TAT	GAA	ACT	CTT	AGA	GTC	AAT	CAG	CCC	480
	Asp	Gly	Gly	Thr	Tyr	Asp	Ile	Tyr	Glu	Thr	Leu	Arg	Val	Asn	Gln	Pro	
	145					150					155					160	
5	TCC	ATT	AAG	GGG	ATT	GCC	ACA	TTT	AAA	CAA	TAT	TGG	AGT	GTC	CGA	AGA	528
	Ser	Ile	Lys	Gly	Ile	Ala	Thr	Phe	Lys	Gln	Tyr	Trp	Ser	Val	Arg	Arg	
					165					170					175		
10	TCG	AAA	CGC	ACG	AGT	GGC	ACA	ATT	TCT	GTC	AGC	AAC	CAC	TTT	AGA	GCG	576
	Ser	Lys	Arg	Thr	Ser	Gly	Thr	Ile	Ser	Val	Ser	Asn	His	Phe	Arg	Ala	
					180					185				190			
15	TGG	GAA	AAC	TTA	GGG	ATG	AAC	ATG	GGG	AAA	ATG	TAT	GAA	GTC	GCG	CTT	624
	Trp	Glu	Asn	Leu	Gly	Met	Asn	Met	Gly	Lys	Met	Tyr	Glu	Val	Ala	Leu	
			195					200					205				
	ACT	GTA	GAA	GGC	TAT	CAA	AGT	AGC	GGA	AGT	GCT	AAT	GTA	TAT	AGC	AAT	672
	Thr	Val	Glu	Gly	Tyr	Gln	Ser	Ser	Gly	Ser	Ala	Asn	Val	Tyr	Ser	Asn	
		210					215					220					
20	ACA	CTA	AGA	ATT	AAC	GGT	AAC	CCT	CTC	TCA	ACT	ATT	AGT	AAT	GAC	AAG	720
	Thr	Leu	Arg	Ile	Asn	Gly	Asn	Pro	Leu	Ser	Thr	Ile	Ser	Asn	Asp	Lys	
		225				230					235					240	
25	AGC	ATA	ACT	CTA	GAT	AAA	AAC	AAT	TAA								747
	Ser	Ile	Thr	Leu	Asp	Lys	Asn	Asn	*								
					245												
30	(2) INFORMATION FOR SEQ ID NO: 2:																
	(i) SEQUENCE CHARACTERISTICS:																
	(A) LENGTH: 249 amino acids																
	(B) TYPE: amino acid																
	(D) TOPOLOGY: linear																
35	(ii) MOLECULE TYPE: protein																
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:																
	Met	Arg	Gln	Lys	Lys	Leu	Thr	Phe	Ile	Leu	Ala	Phe	Leu	Val	Cys	Phe	
	1				5					10					15		
40	Ala	Leu	Thr	Leu	Pro	Ala	Glu	Ile	Ile	Gln	Ala	Gln	Ile	Val	Thr	Asp	
				20					25					30			
45	Asn	Ser	Ile	Gly	Asn	His	Asp	Gly	Tyr	Asp	Tyr	Glu	Phe	Trp	Lys	Asp	
			35					40					45				
	Ser	Gly	Gly	Ser	Gly	Thr	Met	Ile	Leu	Asn	His	Gly	Gly	Thr	Phe	Ser	
		50					55					60					
50	Ala	Gln	Trp	Asn	Asn	Val	Asn	Asn	Ile	Leu	Phe	Arg	Lys	Gly	Lys	Lys	
		65				70					75					80	
	Phe	Asn	Glu	Thr	Gln	Thr	His	Gln	Gln	Val	Gly	Asn	Met	Ser	Ile	Asn	
					85					90					95		
55	Tyr	Gly	Ala	Asn	Phe	Gln	Pro	Asn	Gly	Asn	Ala	Tyr	Leu	Cys	Val	Tyr	
			100						105					110			
	Gly	Trp	Thr	Val	Asp	Pro	Leu	Val	Glu	Tyr	Tyr	Ile	Val	Asp	Ser	Trp	
60			115					120					125				
	Gly	Asn	Trp	Arg	Pro	Pro	Gly	Ala	Thr	Pro	Lys	Gly	Thr	Ile	Thr	Val	
		130					135					140					
65	Asp	Gly	Gly	Thr	Tyr	Asp	Ile	Tyr	Glu	Thr	Leu	Arg	Val	Asn	Gln	Pro	
		145				150					155					160	
	Ser	Ile	Lys	Gly	Ile	Ala	Thr	Phe	Lys	Gln	Tyr	Trp	Ser	Val	Arg	Arg	
					165					170					175		

SUBSTITUTE SHEET (RULE 26)

33

Ser Lys Arg Thr Ser Gly Thr Ile Ser Val Ser Asn His Phe Arg Ala
 180 185 190
 5 Trp Glu Asn Leu Gly Met Asn Met Gly Lys Met Tyr Glu Val Ala Leu
 195 200 205
 Thr Val Glu Gly Tyr Gln Ser Ser Gly Ser Ala Asn Val Tyr Ser Asn
 210 215 220
 10 Thr Leu Arg Ile Asn Gly Asn Pro Leu Ser Thr Ile Ser Asn Asp Lys
 225 230 235 240
 15 Ser Ile Thr Leu Asp Lys Asn Asn *
 245

(2) INFORMATION FOR SEQ ID NO: 3:
 (i) SEQUENCE CHARACTERISTICS:
 20 (A) LENGTH: 8 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: other nucleic acid
 25 (A) DESCRIPTION: /desc = "Conserved region"
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:
 Asp Gly Gly Thr Tyr Asp Ile Tyr
 1 5
 30

(2) INFORMATION FOR SEQ ID NO: 4:
 (i) SEQUENCE CHARACTERISTICS:
 35 (A) LENGTH: 7 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: other nucleic acid
 40 (A) DESCRIPTION: /desc = "Conserved region"
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

Glu Gly Tyr Gln Ser Ser Gly
 1 5
 45

(2) INFORMATION FOR SEQ ID NO: 5:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 29 base pairs
 50 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: other nucleic acid
 (A) DESCRIPTION: /desc = "Primer e"
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

55 GCGAATTCAT GAGACAAAAG AAATTGACG

29

(2) INFORMATION FOR SEQ ID NO: 6:
 (i) SEQUENCE CHARACTERISTICS:
 60 (A) LENGTH: 22 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: other nucleic acid
 65 (A) DESCRIPTION: /desc = "Primer arc "
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

AACAGTGATG GTTCCCTTAG GC

22

SUBSTITUTE SHEET (RULE 26)

34

- (2) INFORMATION FOR SEQ ID NO: 7:
 (i) SEQUENCE CHARACTERISTICS:
 5 (A) LENGTH: 31 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: other nucleic acid
 10 (A) DESCRIPTION: /desc = "Primer f "
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:
 CTAGAGTCGA CTTAATTGTT TTTATCTAGA G 31
- 15 (2) INFORMATION FOR SEQ ID NO: 8:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 22 base pairs
 (B) TYPE: nucleic acid
 20 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: other nucleic acid
 (A) DESCRIPTION: /desc = "Primer drc "
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 8:
 25 AACAGTGATG GTTCCCTTAG GC 22
- (2) INFORMATION FOR SEQ ID NO: 9:
 30 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 42 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: other nucleic acid
 35 (A) DESCRIPTION: /desc = "Primer ab "
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 9:
 GCCTAAGGGA ACCATCACTG TTGAYGGXGG XACXTAYGAY AT 42
- 40 (Y=C or T, X= 25% A and 75% Inosin)
- (2) INFORMATION FOR SEQ ID NO: 10:
 45 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 22 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: other nucleic acid
 50 (A) DESCRIPTION: /desc = "Primer cd "
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 10:
 AATGCTATAT ACATTAGCAC TTCCXSWXSW YTGGTAXCCY TC 42
- 55 (S=G or C, W=A or T, Y=C or T, X= 25% A and 75% Inosin)
- (2) INFORMATION FOR SEQ ID NO: 11:
 60 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 747 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: hybrid DNA
 65 (ix) FEATURE:
 (A) NAME/KEY: CDS
 (B) LOCATION: 1..747
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 11:

SUBSTITUTE SHEET (RULE 26)

35

	ATG AGA CAA AAG AAA TTG ACG TTC ATT TTA GCC TTT TTA GTT TGT TTT	48
	Met Arg Gln Lys Lys Leu Thr Phe Ile Leu Ala Phe Leu Val Cys Phe	
	1 5 10 15	
5	GCA CTA ACC TTA CCT GCA GAA ATA ATT CAG GCA CAA ATC GTC ACC GAC	96
	Ala Leu Thr Leu Pro Ala Glu Ile Ile Gln Ala Gln Ile Val Thr Asp	
	20 25 30	
10	AAT TCC ATT GGC AAC CAC GAT GGC TAT GAT TAT GAA TTT TGG AAA GAT	144
	Asn Ser Ile Gly Asn His Asp Gly Tyr Asp Tyr Glu Phe Trp Lys Asp	
	35 40 45	
15	AGC GGT GGC TCT GGG ACA ATG ATT CTC AAT CAT GGC GGT ACG TTC AGT	192
	Ser Gly Gly Ser Gly Thr Met Ile Leu Asn His Gly Gly Thr Phe Ser	
	50 55 60	
20	GCC CAA TGG AAC AAT GTT AAC AAC ATA TTA TTC CGT AAA GGT AAA AAA	240
	Ala Gln Trp Asn Asn Val Asn Asn Ile Leu Phe Arg Lys Gly Lys Lys	
	65 70 75 80	
25	TTC AAT GAA ACA CAA ACA CAC CAA CAA GTT GGT AAC ATG TCC ATA AAC	288
	Phe Asn Glu Thr Gln Thr His Gln Gln Val Gly Asn Met Ser Ile Asn	
	85 90 95	
30	TAT GGC GCA AAC TTC CAG CCA AAC GGA AAT GCG TAT TTA TGC GTC TAT	336
	Tyr Gly Ala Asn Phe Gln Pro Asn Gly Asn Ala Tyr Leu Cys Val Tyr	
	100 105 110	
35	GGT TGG ACT GTT GAC CCT CTT GTC GAA TAT TAT ATT GTC GAT AGT TGG	384
	Gly Trp Thr Val Asp Pro Leu Val Glu Tyr Tyr Ile Val Asp Ser Trp	
	115 120 125	
40	GGC AAC TGG CGT CCA CCA GGG GCA ACG CCT AAG GGA ACC ATC ACT GTT	432
	Gly Asn Trp Arg Pro Pro Gly Ala Thr Pro Lys Gly Thr Ile Thr Val	
	130 135 140	
45	GAC GGG GGG ACG TAT GAT ATC TAC AAG CAC CAA CAG GTC AAT CAG CCA	480
	Asp Gly Gly Thr Tyr Asp Ile Tyr Lys His Gln Gln Val Asn Gln Pro	
	145 150 155 160	
50	TCT ATT CAG GGC ACC GCC ACC TTC AAT CAG TAC TGG TCG ATT CGA CAG	528
	Ser Ile Gln Gly Thr Ala Thr Phe Asn Gln Tyr Trp Ser Ile Arg Gln	
	165 170 175	
55	AGC AAG CGG ACC AGC GGC ACT GTC ACT ACG GCA AAC CAC TTT AAT GCC	576
	Ser Lys Arg Thr Ser Gly Thr Val Thr Thr Ala Asn His Phe Asn Ala	
	180 185 190	
60	TGC CCT GCT CTT GGC ATG AAT ATG GGT GCA TTC AAT TAC CAG ATC CTC	624
	Trp Ala Ala Leu Gly Met Asn Met Gly Ala Phe Asn Tyr Gln Ile Leu	
	195 200 205	
65	GTT ACT GAG GGC TAC CAA TCT ACC GGA AGT GCT AAT GTA TAT AGC AAT	672
	Val Thr Glu Gly Tyr Gln Ser Thr Gly Ser Ala Asn Val Tyr Ser Asn	
	210 215 220	
70	ACA CTA AGA ATT AAC GGT AAC CCT CTC TCA ACT ATT ACT AAT GAC AAG	720
	Thr Leu Arg Ile Asn Gly Asn Pro Leu Ser Thr Ile Ser Asn Asp Lys	
	225 230 235 240	
75	AGC ATA ACT CTA GAT AAA AAC AAT TAA	747
	Ser Ile Thr Leu Asp Lys Asn Asn *	
	245	
80	(2) INFORMATION FOR SEQ ID NO: 12:	
	(i) SEQUENCE CHARACTERISTICS:	
	(A) LENGTH: 249 amino acids	
	(B) TYPE: amino acid	

SUBSTITUTE SHEET (RULE 26)

36

(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: protein
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 12:

5 Met Arg Gln Lys Lys Leu Thr Phe Ile Leu Ala Phe Leu Val Cys Phe
1 5 10 15
Ala Leu Thr Leu Pro Ala Glu Ile Ile Gln Ala Gln Ile Val Thr Asp
20 25 30
10 Asn Ser Ile Gly Asn His Asp Gly Tyr Asp Tyr Glu Phe Trp Lys Asp
35 40 45
15 Ser Gly Gly Ser Gly Thr Met Ile Leu Asn His Gly Gly Thr Phe Ser
50 55 60
Ala Gln Trp Asn Asn Val Asn Asn Ile Leu Phe Arg Lys Gly Lys Lys
65 70 75 80
20 Phe Asn Glu Thr Gln Thr His Gln Gln Val Gly Asn Met Ser Ile Asn
85 90 95
Tyr Gly Ala Asn Phe Gln Pro Asn Gly Asn Ala Tyr Leu Cys Val Tyr
100 105 110
25 Gly Trp Thr Val Asp Pro Leu Val Glu Tyr Tyr Ile Val Asp Ser Trp
115 120 125
30 Gly Asn Trp Arg Pro Pro Gly Ala Thr Pro Lys Gly Thr Ile Thr Val
130 135 140
Asp Gly Gly Thr Tyr Asp Ile Tyr Lys His Gln Gln Val Asn Gln Pro
145 150 155 160
35 Ser Ile Gln Gly Thr Ala Thr Phe Asn Gln Tyr Trp Ser Ile Arg Gln
165 170 175
Ser Lys Arg Thr Ser Gly Thr Val Thr Thr Ala Asn His Phe Asn Ala
180 185 190
40 Trp Ala Ala Leu Gly Met Asn Met Gly Ala Phe Asn Tyr Gln Ile Leu
195 200 205
Val Thr Glu Gly Tyr Gln Ser Thr Gly Ser Ala Asn Val Tyr Ser Asn
210 215 220
45 Thr Leu Arg Ile Asn Gly Asn Pro Leu Ser Thr Ile Ser Asn Asp Lys
225 230 235 240
50 Ser Ile Thr Leu Asp Lys Asn Asn *
245

(2) INFORMATION FOR SEQ ID NO: 13:
55 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 409 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
60 (ii) MOLECULE TYPE: Hybrid DNA
(vi) SCIENTIFIC NAME: NS1/9
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 13:

GAATTCGGCT	TGGGTGGAAT	CTGGGGAACA	CGTTGGATGC	TACCGGAGAC	TGGATCAAAG	60
GGCCGTCCGT	GAGCGCCTAC	GAGACCGCCT	GGGGCAATCC	CGTCACCACC	AAGGCTATGT	120
TCGACGGCAT	CAAAGCGTCC	GGCTTCAACT	TTGTTGCGAT	TCCCGTGGCG	TGGTCCAACA	180
TTGATGGGCC	GGACTATACC	ATTAACCCGG	CGTTGATGGC	GAGAGTCGAG	AAGTGGTGAA	240
TTACGGTCTG	CGCGACAACA	TGTATGTCAT	GATCAACATC	CACTGGGACG	CGGCTGGATC	300
ACTAAATTCC	CACCAACTAC	GACGAAAGCA	TGAAGAAGTA	TAAGCGGCTC	TGGAGCCAGA	360

SUBSTITUTE SHEET (RULE 26)

TCGCCGACCA TTTCAAAGCT ACTCCGACCA CCTCATCTTC GAAAAGCCG

409

- (2) INFORMATION FOR SEQ ID NO: 14:
- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 408 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
- (vi) SCIENTIFIC NAME: NS1/12
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 14:
- | | | | | | | |
|------------|------------|------------|------------|------------|------------|-----|
| AATTCGGCTT | GGGTGGAATC | TGGGGAACAC | TCTGGAAGCC | TGCGGCGGGA | TCAAATGCAG | 60 |
| 15 | TTCCGTGCGC | GATTTCGAGA | CGGCTTGGGG | CAACCCCGTC | ACGACCAAGG | 120 |
| | CGGCGTCAAG | GCGGCCGGCT | TCAGGTCCAT | ACGCATCCCC | GTCGCCTGGT | 180 |
| | GGGACCTAAG | CCCGACTACA | CTATCAATAA | GAAGCTGATG | GCACGAGTCG | 240 |
| | CCGGTACGGC | CTCGACAACG | ACATGTACGT | CATCATCAAC | ATTCACTGGG | 300 |
| | ATCCACCCCT | TCTCCACCGA | CTACACGAA | ATGCATGARG | AATTACAAGG | 360 |
| 20 | CCAGGTAGCC | GACCATTCA | AGGGTACTC | CGACCACCTC | ATCTTCGA | 408 |

- (2) INFORMATION FOR SEQ ID NO: 15:
- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 416 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
- (vi) SCIENTIFIC NAME: KN1/9
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 15:
- | | | | | | | |
|------------|------------|------------|------------|------------|------------|-----|
| AATTCGGCTT | CTCGAAGATG | AGGTGGTCGG | AGTAGCCTTT | GAAATGGTCC | GCGATCTGGC | 60 |
| TCCAGACCGC | CTTATACTTC | TTCATGCTTT | CGTCGTAGTT | GGTGGGGAAT | TTAGTGATCC | 120 |
| 35 | AGCCGCGGTC | CCAGTGGATG | TTGATCATGA | CATACATGTT | GTCGGCCAGA | 180 |
| | CCACTTCCTC | GACTCTCGCC | ATCAACGCGC | GGTTAATGGT | ATAGTCCGGG | 240 |
| | TGGACCACGC | CACGGGAATG | CGAACAAAGT | TGAAGCCGGA | CGCTTTGATG | 300 |
| | TAGCCTTGGT | GGTGACGGGA | TTGCCCCAGG | CGGTCTCGTA | GGCGCTCAGC | 360 |
| 40 | GATCCAGTC | TCCGGTAGCA | TCCAACGTGT | TCCCCARATT | CCACCCAAGC | 416 |

- (2) INFORMATION FOR SEQ ID NO: 16:
- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 490 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
- (vi) SCIENTIFIC NAME: KM1/2
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 16:
- | | | | | | | |
|------------|------------|------------|------------|------------|------------|-----|
| AATTCGGCTT | GTTCCGCAAG | CGTCAAAGGG | GATGTGATGT | ACCAGATCAA | GGCAAAGCTC | 60 |
| GGTCTGAAAT | AAAACTAGTC | AAAACTAGCC | AAAACTAGTC | AGGCTAGTCA | GAACCACTTA | 120 |
| GCACAATCGT | AAAACTAAA | AGTATGAGCG | ACGGCAATTT | CAACCGCGCC | CTCCTGCCGA | 180 |
| 55 | AGAACGAACT | CTCTGCAGGA | CTCAGGGCTG | GCAAAGCACA | GATGCGCACC | 240 |
| | CAGGCGTTGG | AGACTGTACT | CGACNAATAC | TTCCCCTCTG | CCGACATGTC | 300 |
| | GCAATCCACG | AACGATCCTC | CAACTCTTAC | AACAGTAGGA | CAAAGGTGAA | 360 |
| | TTATGCTTCC | TGAATTNTCA | TTAACACNAT | GCCTGTGTGG | CACCCATCCG | 420 |
| | GGTGTTCACC | AGGGCATCCT | TTACTCATCC | CACAGGTTAA | GCAANTGGCC | 480 |
| 60 | GTCCGGCTTC | | | | AAANAACACC | 490 |

- (2) INFORMATION FOR SEQ ID NO: 17:
- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 492 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
- (vi) SCIENTIFIC NAME: KN2/2

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 17:

AATTCGGCTT GTTGTGCGG CCGGTGGTGC GGACCACGTC AATAAAAGTC TGGTTGTAAG 60
 AATTCTGCAC AGCCAGATTC TCAGGCTCGG GCTTGCCCCA GTTATCGCGC AGGTGAACCT 120
 5 CGTTAGTACC AGCAAAGGCT ACGCGGTAGT CGTAGTTGGC AAACCTCGCTG GCGATATTCA 180
 GCCACAGCAG GCGGAGTTTC TGGTTGTTCT CGTCCTTGTA CTGATAGGTA GGACRACCCT 240
 CCAGCCACTT GTCGTGATGC GTATTGATGA TGACTTTTAG GTCATTCTCG AAGCACCARC 300
 CCACAACCTC TTTGATACGT GCCAGCCAAG CCTTGTCAT GCTCATGGCA ACGGGATTGG 360
 TGATGTTGCA CTGCCACCGG AMSSGAATGC GGATGGCGTT RAAAC: TGCA TCCTTGACTG 420
 10 CTTGATAAC TTTTTTGTTA CAACGGGATT GCCCATGCC GTCTCACCCT TAATACTGTT 480
 CTCATACATC CG 492

(2) INFORMATION FOR SEQ ID NO: 18:

- 15 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 574 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 20 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: KM2/5
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 18:

AATTCGGCTT GTTGTGCGG CCGGTGGTAC GGATGGTGT CACCACCAAC TGGTTCCACT 60
 25 CGTTGAGGGT TTTATACTGC TTACCGCCAT CCGTACGGTT TCGGCCCCAT CCCCAGCCGC 120
 CGTCCTGAAT CTCGTTGAAC GACTCGAATA TGAGGAATTC GCCCTTGTC TGAAGGCTT 180
 CGGCAATCTG TTTCCANGTT TTCTCAATAC GGTCTTGAT GTTGCTGTTG GTCGTTGAAT 240
 TGTGGGCAGC GCCCTTAATG TCAACCAGTA CTCATCGTGA TGCATGTTCA GGATNACNTT 300
 CAGTCCGGCA CTTCCGCCCA CTCCACATTC TGCCTGACTT CTGCTATGTA TTTAGCATCT 360
 30 ATCCCCATTC CAAATGTTTC TGGTANTTGC CCATGTTACC CGANACTTAN GTGCTGGCAC 420
 AACGTTTTTA NGTTTGTTAA AAACCGCAAA GGCTTGGCAT TTCCAATATC CCANTGGGGA 480
 ACCNAACNTC NCACCCNGCC GGTACAAATG GTNCCCNTT TCCCCAACC CAAATCCNCC 540
 NCNGGGGGCC GTTACNATTG NATCNAACCG GTAC 574

(2) INFORMATION FOR SEQ ID NO: 19:

- 35 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 520 base pairs
 40 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: KM2/6
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 19:

AATTCGGCTT GTTGTGCGG CCGGTGGTTC TCACGGTGGT GACGAAGCTC TGAGCATANC 60
 TGTGATGGC GTTGTAGGCC GATGTGGCTA TGGCTTCGTT GTACCTGCCG GTAGCGGCAA 120
 AGGATGCGAA ACACCAGGAG CTCAAGGGAT CCAGCATCTC GTTGAAGCTC TGAAGAGCA 180
 AGCGCTGTCC GCAGTCCCGG AATTCTGTG CTATCTGCTG CCACAGACGT TCATANCGGG 240
 50 AGCGGTTTAN CGCGTATTG TCCTCGGANG CCTTGATCCA CNACTTGAAA CNANTTGCTG 300
 TCTGCGCCCG TGTCGTGGTG AACGTTGAAT NATGCAGTAC AAGCCCTGGT CTAGGANACT 360
 ATCACCATT CATGCACGCG GGCCATCCAC GCCNCATCCA CNTTGCCGGC GCTGTCCATN 420
 TTGTTATACC ACTTCATGCC CCACGGATGG CACCAAACCC GGATCTTTNT CNTCCTGAAN 480
 55 AACAANGGGT GGTGGGATAT TAACCAACA GGTCCGAAGA 520

(2) INFORMATION FOR SEQ ID NO: 20:

- 60 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 194 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: KM3/2
 65 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 20:

AATTCGGCTT GAGCACCTGA TTTTGAGGG CTACAACGAG ATGCTCGACA AGTATGACTC 60
 CTGGTGTGTT GCCACCTTCG GACGCTCGGC AGGCTATAAC GCTACAGACG CCGCCGATGC 120
 CTATAAAGCC ATCAACAACT ATGCCAGAG CTTCGTCAAC GCCGTACGCA CCACCGGCGG 180

CAACAACAAG CCG

194

- (2) INFORMATION FOR SEQ ID NO: 21:
- 5 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 160 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
- 10 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: KM3/8
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 21:
- | | | | | | | |
|---------------|------------|------------|------------|-------------|------------|-----|
| AATTCGGCTT | GAGCACTTGA | TTTTCGAGGC | CTACAACGAG | ATGCTCGATG | CCCAGAGCTC | 60 |
| 15 GTGGAACCTT | GCCCAGACCA | GCACAGCCTA | TGATGCTATC | AACAACATATG | CCCAAAGCTT | 120 |
| CGTCAACATT | GTTCGTACCA | GCGGCGGCAA | CAACAAGCCG | | 160 | |
- (2) INFORMATION FOR SEQ ID NO: 22:
- 20 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 193 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
- 25 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: KM3/9
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 22:
- | | | | | | | |
|---------------|------------|------------|------------|------------|-------------|-----|
| AATTCGGCTT | GAGCATTTGA | TCTTCGAGAG | TTACAACGAG | ATGCTCGATA | CGGAAGATTC | 60 |
| 30 CTGGTGCTTC | GCCTCGTTTG | CAGCGCAGGG | CAGTTACAAT | GCCACCATCG | CGCGTTCCGGC | 120 |
| CTACAACGGC | ATTAATAGCT | ATGCGCAGAC | TTTCGTCAAC | ACCGTACGTA | CCACCGGCGG | 180 |
| CAACAACAAG | CCG | | | | 193 | |
- (2) INFORMATION FOR SEQ ID NO: 23:
- 35 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 166 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
- 40 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: KM4/1
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 23:
- | | | | | | | |
|---------------|------------|------------|------------|------------|------------|-----|
| AATTCGGCTT | CAYACGCTGG | TGTGGCACTC | TCAGATCGGT | CGTTGGATGA | CTGCCGAGGG | 60 |
| 45 TACAACCAAG | GAGCAGTTCT | ATGCTCGTAT | GAAGAACCAT | ATCCAGGCTA | TCGTTACTCG | 120 |
| TTACAAGCAT | GTGGTGTACT | GCTGGGACGT | CGTCAACGAG | AAGCCG | 166 | |
- (2) INFORMATION FOR SEQ ID NO: 24:
- 50 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 178 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
- 55 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: KM4/2
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 24:
- | | | | | | | |
|---------------|------------|------------|------------|------------|------------|-----|
| AATTCGGCTT | CTCGTTAACG | ACGTCCCAGG | CATCGATCTT | ACCGCAGAAA | TGGCCGGCTA | 60 |
| 60 CCGTCTCTAT | GTAAGTGGC | ATGGTCTCAA | CCATCTCATC | GTGGCTCTTG | GGAGTGCCGT | 120 |
| CAGCGTGGTT | GAAAAAGAAA | TCGGGAGTCT | GATTGTGCCA | CACCAGCGTA | TGAAGCCG | 178 |
- (2) INFORMATION FOR SEQ ID NO: 25:
- 65 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 181 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

40

- (ii) MOLECULE TYPE: Hybrid DNA
- (vi) SCIENTIFIC NAME: KM4/4
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 25:

5 AATTCGGCTT CAYACGCTGG TGTGGCACTC GCAGGCACCC GACTGGTGGT TTACCAACGG 60
 CTATGCTGCC AGCCCTGTCT CAAAGGAAGT GCTGAAAGAG CGGCTCATCA AGCATATTAA 120
 GACCGTTGTT GGCCATTCA AGGGCCAAGT CTTGGCTGG GACGTCGTCA ACGARAAGCC 180
 G 181

10

- (2) INFORMATION FOR SEQ ID NO: 26:
- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 199 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
- (vi) SCIENTIFIC NAME: KM4/7
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 26:

15 AATTCGGCTT CATACTGTGG TGTGGCACAA TCAGACGCCG GCCTGGTTCT TCCGCAGGGG 60
 CTACAACGAG AACCTGCCTC TGGCGGACCG CGAGACCATG CTGGCGAGGC TGGAGAGCTA 120
 TATCCGCGGT GTGCTGACCT ATGTGCAGGA GAATTATCCC GGGATCGTCT ACGCCTGGGA 180
 25 CGTCGTCAAC GAGAAGCCG 199

- (2) INFORMATION FOR SEQ ID NO: 27:
- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 185 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
- (vi) SCIENTIFIC NAME: KM4/8
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 27:

30 AATTCGGCTT GGCACGGACA GACGCCGCGAG TGGTTCTTCT ACGAGAACTA TAATACTTCA 60
 GGAAAACCTG CAAGCAGGGA AACGATGCTG GCAAGAATGG GAACTATAT TAANGGCGTG 120
 40 CTTGGCTTCG TGCAGGACAA TTATCCCGGC GTCATCTATG CGTGGGACGT TGTCACGAG 180
 AACCG 185

- (2) INFORMATION FOR SEQ ID NO: 28:
- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 208 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
- (vi) SCIENTIFIC NAME: KM4/9
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 28:

45 ATCTGCAGAA ATTCGGCTTC TCGTTAACGA CGTCCCATGC ATAGATGACA CCCGGATATT 60
 CACTCTGGAT AAAACCAAGC ACACCCITTA TATAATTTT AAGTCTGGCA AGCATGGTCT 120
 CTCTGTCGGT ATAGGGAAAT GACTCGTTAT AGTGCTCACA GAAAAACCAC TTCGGTGTCT 180
 GATTGTGCCA CACCAGCGTA TGAAGCCC 208

- (2) INFORMATION FOR SEQ ID NO: 29:
- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 310 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
- (vi) SCIENTIFIC NAME: KM5/1
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 29:

65

SUBSTITUTE SHEET (RULE 26)

41

AATTCGGCTT GTTGTAGTCG TTGTAGTACA GCTTGCACTT TGAAGGAGCG TACTTTCTTG 60
 CATATGTGAA CGCTTTCTCA ATAAATGCGT TGCTGCCGTA AACCTGTACC CAAGGGANAA 120
 GCGCCGTTGC CGTACCCGGA ACTCTTGCTC CGCCGTTGTT ACGTGTTCTG TTGGAGTCAC 180
 ANAAAATACA CTCGTTGCAG ACATCTAAAG CTAAAGGTT AATCCGGGAT ACTGTGACTG 240
 5 ATAGGCCGAA CATATCTTGA AGTTACCTTC CAGTCCNGGT CCATACGGAA TGCTACCAGC 300
 TTCGCCGTCC 310

- (2) INFORMATION FOR SEQ ID NO: 30:
 10 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 384 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 15 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: KM5/2
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 30:

AATTCGGCTT GTTGTANTNG TTGWWGAAGA NGTGGCAGNT TGCCGGTGCC GCATCATGGG 60
 20 CATATTCAAA TGCCTTTGCA ATGAAGCTGT TGTCACCGTA AACCTGCACC CACGGGGACT 120
 TGCCGTGATT GTAACCCGGC TCACGGGGCGC CGCCTGCACC ACGCGTACGC GCATCGCTGT 180
 CGGAGATACA CTCGTTGCAG ACGTCGTARG CGTANARGTT CAGCGTCNGA TAGTTGTTCT 240
 TGTACATTGC AAMCATATTG TCAATGTANC YCTTGANGCG CTGTTTCATG ACAGTGGANT 300
 TCACCCACTG ACCGCGCTCC TGGAAAGTTA TCCTTGAAAN AACCAGANCG GARTCTGGRA 360
 25 GTGCCACNCC ANCGTRTGAA GCCG 384

- (2) INFORMATION FOR SEQ ID NO: 31:
 (i) SEQUENCE CHARACTERISTICS:
 30 (A) LENGTH: 354 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: KM5/4
 35 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 31:

AATTCGGCTT CATACGTTGG TGTGGCACAA TCAGACGCCC GATGTTT TTAAGGAAAA 60
 CTGGGAAAT GACTGGAACG CGCCTGCCGC CCCCAGAGAA ATCCTGCTCG CCCGCTGGA 120
 40 AAACATATC CGGGATGTCA TCGGGCATGT GAATACCTGT TTCCCCGGTG TGGTCTACAC 180
 CTGGGATGTG GTGAACGAAG CCATCGAACC GGGGCAGGGC GGTCCCGGCC TGTCCGGAA 240
 CCGCAATCCC TGGTTTGCTT TCACAGGCCA NGATTTCTG CCGGCTGCCT TCCGGGCCCC 300
 CGCGAAAACN AAGTCCCGGG ACAGAACCTG TGCTACAACG ACTACAACAA GCCG 354

- 45 (2) INFORMATION FOR SEQ ID NO: 32:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 374 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 50 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: KM5/5
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 32:

55 AATTCGGCTT CATACGCTGG TGTGGCACAG CCAGACTCCT GACTGGTCT TCAAGGAGAA 60
 CTCAGCTCA AACGGTCAGC TCGTATCAAA GGATATAATG AATCAGCGTA TCGAAAACATA 120
 CATCAAGAAC GTATTCACAA TGCTCAATGC AGAGTATCCT ACAGTTCAGT TCTATGCTTA 180
 CGATGTAGCT AACGAGTGTG TGGCTGACAG CAGAAACGGC GGTCTCAGAC CGGCTGGCAT 240
 GAATCAGCAG AACGGCGAAT CCCCATGGAA TCTTATCTAC GGCGACAACA GCTACCTCGA 300
 60 TGTANCATTG AAGGCTGCTA AGAAATTATG CTCCTGCTGG CTGCNAACCT TTCTTCAACG 360
 ACTACAACAA GCCG 374

- (2) INFORMATION FOR SEQ ID NO: 33:
 65 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 376 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

SUBSTITUTE SHEET (RULE 26)

42

- (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: KM5/6
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 33:

5 AATTCGGCTT CATACGCTGG TGTGGCACAG CCAGACTCCC GAGTGGTTCT TCAAGGAGGA 60
 CTCGACGAG AAGAAGGATT ACGTTTCTCC CGAAAAGATG AAGAAGCGTA TGGAGAATA 120
 CATCAAGAGC TTCTTCACAA CACTTACAGA GCTCTATCCC GACGTTGACT TCTATGCCTG 180
 CGACGTTGTA AACGANGCAT GGACAGACGA CGGAAAGCCC CGTGAGGCAG GTCAGTGTTC 240
 ACAGTCCAAC AACTACGGCG CTTCCGACTG GGTGCTGTA TTCGGCGACA ACTCATTCT 300
 10 CGACTACGCT TTCGAGTATG CAAGAAAGTA TGCTCCCGAN GGCTGCAAGC TCTACTACAA 360
 CGACTACAAC AAGCCG 376

(2) INFORMATION FOR SEQ ID NO: 34:

- (i) SEQUENCE CHARACTERISTICS:
 15 (A) LENGTH: 166 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: Hybrid DNA
 20 (vi) SCIENTIFIC NAME: NS6/3
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 34:

AATTCGGCTT TGGGATGTGG TGAACGAGGC CTTCAACGAA GACGGTTCAC GGCGCAGCGA 60
 CGTTTCCAG AATGTGCTCG GCAACGGCTA TATCGAGCAG GCATTACAGGA CCGCGCGTGC 120
 25 GGCTGACCCC AATGCCAAAC TGTGCTACAA CGACTACAAC AAGCCG 166

(2) INFORMATION FOR SEQ ID NO: 35:

- (i) SEQUENCE CHARACTERISTICS:
 30 (A) LENGTH: 151 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: NS6/5
 35 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 35:

AATTCGGCTT GTTGTAGTCG TTGTTGAACA GGCGGGTGGT TGGGTCTACC TCATGAGCAA 60
 GTTGATACCA GTGCACAACA GCATCGAGGC CGCCGAGGGC ATCATAAACC TCGTGGTTAT 120
 40 CTACCGGCTC GTTCACCACA TCCCAAAGCC G 151

(2) INFORMATION FOR SEQ ID NO: 36:

- (i) SEQUENCE CHARACTERISTICS:
 45 (A) LENGTH: 166 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: NS6/13
 50 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 36:

AATTCGGCTT GTTGTAGTCG TTGTAGCACA GTTTGGCATT GGGATCTGTA ACCCGTGCAG 60
 CTTGAATCC CTCTTCAATA TAGCTATTGC CAATCAGCCG TTGGAAGATT GAGGCACGCC 120
 GTGAGCCATT GTCTTCGAAG GCCTCATTCA CCACATCCCA AAGCCG 166

(2) INFORMATION FOR SEQ ID NO: 37:

- (i) SEQUENCE CHARACTERISTICS:
 55 (A) LENGTH: 250 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: NS6A/1
 60 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 37:

65 AATTCGGCTT GTTGTAGTCG TTGWTGMAGA GTTTTACATC TTTTGGACCA TATTGCGAG 60
 CCAGACGACA GGCCTGACGG ACGTAGTCGA TATCACCAG ATAGTCCTGC CAGTAGAAAT 120
 TATCGCCGCC CACATCCCAT GTGGCATCTG GATTACCATT AGGATTATAC TTAGCAGACT 180
 GTTGTAATAA GTAGTTGCCT TGTCCGTCAT CACCACCACC AGAGATCGCC TCRTTCACCA 240
 CATCCCAAAG 250

SUBSTITUTE SHEET (RULE 26)

- (2) INFORMATION FOR SEQ ID NO: 38:
- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 247 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
- (vi) SCIENTIFIC NAME: KM6A/4
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 38:
- ```

AATTCGGCTT TGGGAYGTGG TGAAYGAGGC GATAGAGCTT AACGACAAGA CCGAAACCCG 60
ACTTCGTAAT TCATACTGGT ATCAAATAAT CGGTGACGAT TTCATATATT ACGCATTTCG 120
CTATGCATAT GACGCAAGAG AGGAACTGTG CGTTAAATAT GCGGCCGAGT ACGGCATTGA 180
CCCTTCGGAC AAAGAAGCGC TTAAAGCCAT CCGCCCCGCT TTCTGCAACA ACGACTACAA 240
CAAGCCG 247

```
- (2) INFORMATION FOR SEQ ID NO: 39:
- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 238 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
- (vi) SCIENTIFIC NAME: KM6A/5
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 39:
- ```

AATTCGGCTT TGGGATGTGG TGAACGAGGC TATCTCGGGT GGCGACAGTG ACGGCCGACGG 60
TTACTACGAC CTCCAGCATT CCGAGGGGCTA TAAGAACGGC ACTTGGGATG TAGGCGGCCGA 120
TGCCTTCTAC TGGCAGGACT ACATGGGCGA CCTGGATTAC GTRCGTCAGG CTTGCCGACT 180
GGCCCCGAAA TACGGCCCTG AGGATGTGAA GCTYTKCATC AACGACTACA ACAAGCCG 238

```
- (2) INFORMATION FOR SEQ ID NO: 40:
- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 226 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
- (vi) SCIENTIFIC NAME: KM6A/7
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 40:
- ```

AATTCGGCTT GTTGTAGTCG TTGATGCACA ACAGGGCATT GGGGTCTGGCC TCACGGGCAA 60
ACTCGAAAGC TTTGGCAATG AACTCGTCGC CGCAGAGTTT GTAATGACGA CTCTCAGCAT 120
AGGGGCTGGG AGCCTGACCT GGACGGCGTC CGAAACCGCC AAAGCCACCA AAGCCACCAA 180
AGCCGCCACC GTCGGAAATG GCCTCGTTCA CTACATCCCA AAGCCG 226

```
- (2) INFORMATION FOR SEQ ID NO: 41:
- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 205 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
- (vi) SCIENTIFIC NAME: KM6B/1
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 41:
- ```

ATCTGCAGAA ATTCGGCTTT GGGACGTGGT GAACGAGGCT ATGGCCGACG ACGTTCGCCG 60
CTCGCCCTGG AACCCGAATC CGTCGCCCTTA CCGCAACTCG AAACCTCTATC AGTTGTGCCG 120
TGATGAGTTC ATCGCTAAAG CATTCCAATT CGCCCGTGAG GCCGACCCGA ACGCACAATT 180
GTGCATCAAC GACTACAACA AGCCG                                     205

```
- (2) INFORMATION FOR SEQ ID NO: 42:
- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 235 base pairs
- (B) TYPE: nucleic acid

44

(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: Hybrid DNA
(vi) SCIENTIFIC NAME: KM6B/2
5 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 42:

```

AATTCGGCTT GTTGTAGTCG TTGATGAAGA GCTTCATATC CTGTGGACCA TACTTGCGAG 60
CCAGCTTAAC GGCAGTACGA ACATAGTCGA TATCGCCCAG ATAATCCTGC CAGAAGAAGC 120
TCTCGGTTGC AGCCTTTTCT GGATCTTCCT GATCCTTCAG GTGCTGCAA GCATATACGC 180
10 CCTCAGCATC GGCATGTCCG CTTGAGAGTG CCTCGTTCAC CACATCCCAA AGCCG 235

```

(2) INFORMATION FOR SEQ ID NO: 43:
(i) SEQUENCE CHARACTERISTICS:
15 (A) LENGTH: 244 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: Hybrid DNA
(vi) SCIENTIFIC NAME: KM6B/3
20 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 43:

```

AATTCGGCTT GTTGTAGTCG TTGATGAANA GTTTCAGTC TTCCGGGTTG CCCTTGAAGT 60
GCTTGCGCGC ACTCTTAACC GCGGTACGCA CGTATTCCAN GTCGCCCATA TCGTCTGCC 120
AAAAGAAAG CCATTCTGCA CTGAAGTCGG GTCGGTGTG CCGCTACTGT TGTGCTGAAN 180
25 GGGATAATTG CCCTGCCCAT CGTTGCCGCC GCCAGANATA CCTCGTTCAC ACGTCCCAA 240
GCCG 244

```

(2) INFORMATION FOR SEQ ID NO: 44:
(i) SEQUENCE CHARACTERISTICS:
30 (A) LENGTH: 212 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: Hybrid DNA
35 (vi) SCIENTIFIC NAME: KM6B/4
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 44:

```

AAATTCGGCT TGTTGTAGTC GTTGATGTAC AGGACCGGGG CTTTGCCGTA CTTGGCGCAA 60
GCCTCTGTTG CATAGGCGAA TGCAGCATCA ACCCAGTCTT TGGTGCTCGG GTAATAATTG 120
40 CCCAGACAA AGTCGTTGCC AGATGCTCCC TGGGTGCGGA ATGCCCCGCC GGCACCGTCT 180
GCAAAGGTCT CGTTCACCAC GTCCCAAAGC CG 212

```

(2) INFORMATION FOR SEQ ID NO: 45:
(i) SEQUENCE CHARACTERISTICS:
45 (A) LENGTH: 190 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: Hybrid DNA
50 (vi) SCIENTIFIC NAME: KM6B/5
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 45:

```

AATTCGGCTT GTTGTAGTCG TTGTAGAACA GACCTGCATT AGGATCAGCC TCGTGAGCAA 60
ACTGGAATGC CTTGAGGATG AACTCGTCAC CGCAGAGCTG ATAAGCGGTT GACTGACGGA 120
55 ATGACTGCTC GTAAGGAACA TCGGGGTTGT TGCCGTCGCT CATTGCCTCG TTTACCACGT 180
CCCAAAGCCG 190

```

(2) INFORMATION FOR SEQ ID NO: 46:
(i) SEQUENCE CHARACTERISTICS:
60 (A) LENGTH: 234 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: Hybrid DNA
65 (vi) SCIENTIFIC NAME: NS8/1
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 46:

```

AATTCGGCTT GACGGGGGGA CGTAYGAYAT CTACGAGACC ACCCGCTACA ACGAACCCCTC 60
CATCATCGGC ACCGCCACCT TCAACCAGTA CTGGAGCGTG CGCCAGTCCA GGCGCACCGG 120

```

SUBSTITUTE SHEET (RULE 26)

45

CGGCACCATC ACCACCGGCA ACCACTTCGA CGCCTGGGCC AGCCACGGCA TGAACCTGGG 180
CACCTTCAAC TACCAGATCC TGGCCACCGA RGGCTACCAA TSCTSCGGAA GCCG 234

- 5 (2) INFORMATION FOR SEQ ID NO: 47:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 234 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 10 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: NS8/6
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 47:

15 AATTCGGCTT GACGGGGGGA CGTACGACAT CTACGAGCAC CAGCAAGTCA ACCAGCCCTC 60
 CATCCAAGGC ACTGCGACCT TCAACCAAGTA CTGGTCCATC CGCCAGAGCA AGCGTCCAG 120
 CGGCACTGTG ACCACTGCCA ACCACTTCAA TGCTTGGGCC AAGTTGGGAA TGAACCTGGG 180
 CAACTTCAAC TACCAGATTG TTCCACTGA RGGCTACCAG WCCTSCGGAA GCCG 234

- 20 (2) INFORMATION FOR SEQ ID NO: 48:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 234 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 25 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: NS8/11
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 48:

30 AATTCGGCTT GACGGGGGGA CGTATGATAT CTACAAGCAC CAACAGGTCA ATCAGCCATC 60
 TATTCAGGGC ACCGCCACCT TCAATCAGTA CTGGTCCGATT CGACAGAGCA AGCGGACCAG 120
 CGGCACTGTC ACTACGGCAA ACCACTTTAA TGCCTGGGCT GCTCTTGGCA TGAATATGGG 180
 TGCAATCAAT TACCAGATCC TCGTTACTGA GGGCTACCAA TCTACCGGAA GCCG 234

- 35 (2) INFORMATION FOR SEQ ID NO: 49:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 213 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 40 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: NS8/12
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 49:

45 AATTCGGCTT GACGGGGGGA CGTACGACAT TTATGAAACA ACCCGTGTCA ATCAGCCCTC 60
 CATTATCGGG ATCGCAACCT TCAAGCAATA TTGGAGTGTA CGTCAAACGA AACGTACAAG 120
 CGGAACGGTC TCCGTCAGTG CGCATTTTAG AAAATGGGAA AGCTTAGGGA TGCCAATGGG 180
 GAAATGTAT GAAACGGCAT TTAAGTGAAG CCG 213

- 50 (2) INFORMATION FOR SEQ ID NO: 50:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 196 base pairs
 (B) TYPE: nucleic acid
 55 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: Hybrid DNA
 (vi) SCIENTIFIC NAME: KM8A/1
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 50:

60 AATTCGGCTT TGGGACGTGG TGAATGAGGC AATGGCAGAC AATGTTGTC CTAACCCGTG 60
 GAATCCCAAC CCCTCGCCCT ACCGTGACTC CCGCCACTAC AAATGTGCG GCGACGAGTT 120
 CATCGCCAAG GCATTCCAAT TCGCAAGGGA AGCCGACCCG AAGGCACAAT TGTTCACAA 180
 CGACTACAAC AAGCCG 196

65

- (2) INFORMATION FOR SEQ ID NO: 51:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 211 base pairs

SUBSTITUTE SHEET (RULE 26)

46

- (B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
- 5 (ii) MOLECULE TYPE: Hybrid DNA
(vi) SCIENTIFIC NAME: KM8A/3
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 51:
- | | | | | | | |
|---------------|------------|------------|------------|------------|------------|-----|
| AATTCGGCTT | GTTGTAGTCG | TTGATGCACA | GGACCGGGGC | TTTGCCGTAC | TTGGCGCAAG | 60 |
| CCTCTGTTGC | ATAGGCCAAT | GCAGCATCAA | CCCAGTCTTT | GGTGCTCGGG | TAATAATTGC | 120 |
| 10 CCAAAACAAA | GTGCTTGGCA | GATGCTCCCT | GGGTGCGGAA | TGCCCCGCCG | GCACCGTCTG | 180 |
| CAAAGGTCTC | GTTCAACACG | TCCCAAAGCC | G | | 211 | |
- (2) INFORMATION FOR SEQ ID NO: 52:
- 15 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 240 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
20 (vi) SCIENTIFIC NAME: KM8B/7
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 52:
- | | | | | | | |
|---------------|------------|------------|------------|------------|------------|-----|
| AATTCGGCTT | GACGGGGGGA | CGTACGACAT | CTACAAGACC | ACCAGATACG | AACAGCCCTC | 60 |
| TATCGACGGC | ACACAGACCT | TCGACCAGTA | CTGGAGCGTA | AGACAGTCCA | AGCCACAGGG | 120 |
| 25 CGAGGGCAAG | AGATAGAAG | GTACTATCTC | AGTGTCGAAG | CACTTCGATG | CGTGGAAAAA | 180 |
| GTGCGGCCTT | GAGCTCGGAA | ATATGTATGA | AGTANCTCTT | ACTATCGAAG | GGCTAAGCCG | 240 |
- (2) INFORMATION FOR SEQ ID NO: 53:
- 30 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 229 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
35 (vi) SCIENTIFIC NAME: KM8A/9
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 53:
- | | | | | | | |
|--------------|------------|------------|------------|------------|------------|-----|
| AATTCGCCGA | GGTTTGGCAG | CCTTCAATAG | TAAGAGCAGC | TTCATACATT | AATCCTAATT | 60 |
| TCATTCCTTT | GCTTGTCCAA | GCTTTGAAGT | GGTCACCTAC | AGAAATAGTT | CCACTAGTTT | 120 |
| 40 TTTTTCAGT | TCGACACTC | CAGAATTGTT | TAAATGTAGC | AGTACCATCA | ATTGAAGGTT | 180 |
| GATTAATTCT | GTCAAGTGTA | TANATATCAT | ACGTCCCCCC | ATCAAGCCG | 229 | |
- (2) INFORMATION FOR SEQ ID NO: 54:
- 45 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 234 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: Hybrid DNA
50 (vi) SCIENTIFIC NAME: KM8B/10
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 54:
- | | | | | | | |
|---------------|------------|------------|------------|------------|------------|-----|
| AATTCGGCTT | GACGGGGGGA | CGTACGACAT | ATACGAGACT | ACTCGTTACA | ACCAGCCTTC | 60 |
| AATCGAAGGC | AACACTACTT | TCCAGCAGTA | CTGGAGCGTT | CGTACATCCA | AGCGCACCAG | 120 |
| 55 CCGTACCATT | TCCGTATCCG | AGCACTTTAA | GGCTTGGGAA | CGCATGGGTA | TGAGATGCCG | 180 |
| AAACCTTTAT | GAGACTGCTT | TAACTGTTGA | GGGCTACCAN | ACCACCGGAA | GCCG | 234 |
- (2) INFORMATION FOR SEQ ID NO: 55:
- 60 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1060 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: cDNA
65 (iii) HYPOTHETICAL: NO
(vi) ORIGINAL SOURCE:
(A) ORGANISM: Humicola insolens
(B) STRAIN: DSM 1800
(ix) FEATURE:

SUBSTITUTE SHEET (RULE 26)

47

	(A) NAME/KEY: mat_peptide	
	(B) LOCATION: 73..927	
	(ix) FEATURE:	
	(A) NAME/KEY: sig_peptide	
5	(B) LOCATION: 10..72	
	(ix) FEATURE:	
	(A) NAME/KEY: CDS	
	(B) LOCATION: 10..927	
10	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:55:	
	GGATCCAAG ATG CGT TCC TCC CCC CTC CTC CCG TCC GCC GTT GTG GCC	48
	Met Arg Ser Ser Pro Leu Leu Pro Ser Ala Val Val Ala	
	-21 -20 -15 -10	
15	GCC CTG CCG GTG TTG GCC CTT GCC GCT GAT GGC AGG TCC ACC CGC TAC	96
	Ala Leu Pro Val Leu Ala Leu Ala Ala Asp Gly Arg Ser Thr Arg Tyr	
	-5 1 5	
20	TGG GAC TGC TGC AAG CCT TCG TGC GGC TCG GCC AAG AAG GCT CCC GTG	144
	Trp Asp Cys Cys Lys Pro Ser Cys Gly Trp Ala Lys Lys Ala Pro Val	
	10 15 20	
25	AAC CAG CCT GTC TTT TCC TGC AAC GCC AAC TTC CAG CGT ATC ACG GAC	192
	Asn Gln Pro Val Phe Ser Cys Asn Ala Asn Phe Gln Arg Ile Thr Asp	
	25 30 35 40	
30	TTC GAC GCC AAG TCC GGC TGC GAG CCG GGC GGT GTC GCC TAC TCG TGC	240
	Phe Asp Ala Lys Ser Gly Cys Glu Pro Gly Gly Val Ala Tyr Ser Cys	
	45 50 55	
35	GCC GAC CAG ACC CCA TGG GCT GTG AAC GAC GAC TTC GCG CTC GGT TTT	288
	Ala Asp Gln Thr Pro Trp Ala Val Asn Asp Asp Phe Ala Leu Gly Phe	
	60 65 70	
40	GCT GCC ACC TCT ATT GCC GGC AGC AAT GAG GCG GGC TGG TGC TGC GCC	336
	Ala Ala Thr Ser Ile Ala Gly Ser Asn Glu Ala Gly Trp Cys Cys Ala	
	75 80 85	
45	TGC TAC GAG CTC ACC TTC ACA TCC GGT CCT GTT GCT GGC AAG AAG ATG	384
	Cys Tyr Glu Leu Thr Phe Thr Ser Gly Pro Val Ala Gly Lys Lys Met	
	90 95 100	
50	GTC GTC CAG TCC ACC AGC ACT GGC GGT GAT CTT GGC AGC AAC CAC TTC	432
	Val Val Gln Ser Thr Ser Thr Gly Gly Asp Leu Gly Ser Asn His Phe	
	105 110 115 120	
55	GAT CTC AAC ATC CCC GGC GGC GGC GTC GGC ATC TTC GAC GGA TGC ACT	480
	Asp Leu Asn Ile Pro Gly Gly Gly Val Gly Ile Phe Asp Gly Cys Thr	
	125 130 135	
60	CCC CAG TTC GGC GGT CTG CCC GGC CAG CGC TAC GGC GGC ATC TCG TCC	528
	Pro Gln Phe Gly Gly Leu Pro Gly Gln Arg Tyr Gly Gly Ile Ser Ser	
	140 145 150	
65	CGC AAC GAG TGC GAT CGG TTC CCC GAC GCC CTC AAG CCC GGC TGC TAC	576
	Arg Asn Glu Cys Asp Arg Phe Pro Asp Ala Leu Lys Pro Gly Cys Tyr	
	155 160 165	
70	TGG CGC TTC GAC TGG TTC AAG AAC GCC GAC AAT CCG AGC TTC AGC TTC	624
	Trp Arg Phe Asp Trp Phe Lys Asn Ala Asp Asn Pro Ser Phe Ser Phe	
	170 175 180	
75	CGT CAG GTC CAG TGC CCA GCC GAG CTC GTC GCT CGC ACC GGA TGC CGC	672
	Arg Gln Val Gln Cys Pro Ala Glu Leu Val Ala Arg Thr Gly Cys Arg	
	185 190 195 200	
80	CGC AAC GAC GAC GGC AAC TTC CCT GCC GTC CAG ATC CCC TCC AGC AGC	720
	Arg Asn Asp Asp Gly Asn Phe Pro Ala Val Gln Ile Pro Ser Ser Ser	
	205 210 215	

SUBSTITUTE SHEET (RULE 26)

48

	ACC AGC TCT CCG GTC AAC CAG CCT ACC AGC ACC AGC ACC ACG TCC ACC	768
	Thr Ser Ser Pro Val Asn Gln Pro Thr Ser Thr Ser Thr Thr Ser Thr	
	220 225 230	
5	TCC ACC ACC TCG AGC CCG CCA GTC CAG CCT ACG ACT CCC AGC GGC TGC	816
	Ser Thr Thr Ser Ser Pro Pro Val Gln Pro Thr Thr Pro Ser Gly Cys	
	235 240 245	
10	ACT GCT GAG AGG TGG GCT CAG TGC GGC GGC AAT GGC TGG AGC GGC TGC	864
	Thr Ala Glu Arg Trp Ala Gln Cys Gly Gly Asn Gly Trp Ser Gly Cys	
	250 255 260	
15	ACC ACC TGC GTC GCT GGC AGC ACT TGC ACG AAG ATT AAT GAC TGG TAC	912
	Thr Thr Cys Val Ala Gly Ser Thr Cys Thr Lys Ile Asn Asp Trp Tyr	
	265 270 275 280	
20	CAT CAG TGC CTG TAGACGCAGG GCAGCTTGAG GGCCTTACTG GTGGCCGCAA	964
	His Gln Cys Leu	
	285	
	CGAAATGACA CTCCCAATCA CTGTATTAGT TCTTGACAT AATTTCGTCA TCCCTCCAGG	1024
25	GATTGTCACA TAAATGCAAT GAGGAACAAT GAGTAC	1060
	(2) INFORMATION FOR SEQ ID NO:56:	
	(i) SEQUENCE CHARACTERISTICS:	
30	(A) LENGTH: 305 amino acids	
	(B) TYPE: amino acid	
	(D) TOPOLOGY: linear	
	(ii) MOLECULE TYPE: protein	
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO:56:	
35	Met Arg Ser Ser Pro Leu Leu Pro Ser Ala Val Val Ala Ala Leu Pro	
	-21 -20 -15 -10	
	Val Leu Ala Leu Ala Ala Asp Gly Arg Ser Thr Arg Tyr Trp Asp Cys	
	-5 1 5 10	
40	Cys Lys Pro Ser Cys Gly Trp Ala Lys Lys Ala Pro Val Asn Gln Pro	
	15 20 25	
45	Val Phe Ser Cys Asn Ala Asn Phe Gln Arg Ile Thr Asp Phe Asp Ala	
	30 35 40	
	Lys Ser Gly Cys Glu Pro Gly Gly Val Ala Tyr Ser Cys Ala Asp Gln	
	45 50 55	
50	Thr Pro Trp Ala Val Asn Asp Asp Phe Ala Leu Gly Phe Ala Ala Thr	
	60 65 70 75	
	Ser Ile Ala Gly Ser Asn Glu Ala Gly Trp Cys Cys Ala Cys Tyr Glu	
	80 85 90	
55	Leu Thr Phe Thr Ser Gly Pro Val Ala Gly Lys Lys Met Val Val Gln	
	95 100 105	
60	Ser Thr Ser Thr Gly Gly Asp Leu Gly Ser Asn His Phe Asp Leu Asn	
	110 115 120	
	Ile Pro Gly Gly Gly Val Gly Ile Phe Asp Gly Cys Thr Pro Gln Phe	
	125 130 135	
65	Gly Gly Leu Pro Gly Gln Arg Tyr Gly Gly Ile Ser Ser Arg Asn Glu	
	140 145 150 155	
	Cys Asp Arg Phe Pro Asp Ala Leu Lys Pro Gly Cys Tyr Trp Arg Phe	
	160 165 170	

SUBSTITUTE SHEET (RULE 26)

Leu

25

(2) INFORMATION FOR SEQ ID NO: 57:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 9 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: other nucleic acid
 - (A) DESCRIPTION: /desc = "Conserved region"
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 57:

35

Thr Arg Tyr Trp Asp Cys Cys Lys Pro/Thr
1 5

40

(2) INFORMATION FOR SEQ ID NO: 58:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 6 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: other nucleic acid
 (A) DESCRIPTION: /desc = "Conserved region"
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 58:

50

Trp Arg Phe/Tyr Asp Trp Phe
1 5

(2) INFORMATION FOR SEQ ID NO: 59:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 41 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: other nucleic acid
 (A) DESCRIPTION: /desc = "Primer s"
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 59:

GCTGATGGCA GGTCCACIA/CG ITAC/TTGGGAC/T TGC/TTGC/TAAA/GA/C C

41

65

(2) INFORMATION FOR SEQ ID NO: 60:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 29 base pairs
 (B) TYPE: nucleic acid

50

- (C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer as"
5 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 60:
GTGCGCGTTC TTA/GAACCAA/GT CA/GA/TAICG/TCC 29
- (2) INFORMATION FOR SEQ ID NO: 61:
10 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
15 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "forward primer 1"
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 61:
TGGTTC/TAAGA ACGCCGACAA TCCG 24
- (2) INFORMATION FOR SEQ ID NO: 62:
20 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 base pairs
(B) TYPE: nucleic acid
25 (C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "reverse primer 1"
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 62:
30 GCTCTAGAGC CTGCGTCTAC AGGCACTGAT 30
- (2) INFORMATION FOR SEQ ID NO: 63:
35 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 93 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
40 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "forward primer 2"
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 63:
CGGGATCCCA TTTATGATGG TCGCGTGGTG GTCTCTATTT CTGTACGGCC
45 TTCAGGTCGC GGCACCTGCT TTCGCTGCTG ATGGCAGGTC CAC 93
- (2) INFORMATION FOR SEQ ID NO: 64:
50 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 30 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "reverse primer 2"
55 (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 64:
GCTCTAGAGC CTGCGTCTAC AGGCACTGAT 30
- (2) INFORMATION FOR SEQ ID NO: 65:
60 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 922 base pairs
(B) TYPE: nucleic acid
65 (C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: hybrid DNA

SUBSTITUTE SHEET (RULE 26)

(ix) FEATURE:

(A) NAME/KEY: CDS

(B) LOCATION:1..922

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 65:

5	CCA	TTT	ATG	ATG	GTC	GCG	TGG	TGG	TCT	CTA	TTT	CTG	TAC	GGC	CTT	CAG	48
	Pro	Phe	Met	Met	Val	Ala	Trp	Trp	Ser	Leu	Phe	Leu	Tyr	Gly	Leu	Gln	
	1				5				10					15			
10	GTC	GCG	GCA	CCT	GCT	TTC	GCT	GCT	GAT	GGC	AGG	TCC	ACG	CGG	TAC	TGG	96
	Val	Ala	Ala	Pro	Ala	Phe	Ala	Ala	Asp	Gly	Arg	Ser	Thr	Arg	Tyr	Trp	
				20					25					30			
15	GAT	TGC	TGT	AAG	CCG	TCG	TGC	TCG	TGG	CCC	GGC	AAG	GCG	CTC	GTG	AAC	144
	Asp	Cys	Cys	Lys	Pro	Ser	Cys	Ser	Trp	Pro	Gly	Lys	Ala	Leu	Val	Asn	
			35					40					45				
20	CAG	CCC	GTC	TAC	GCC	CGC	AAC	GCA	AAC	TTC	CAG	CGC	ATC	ACC	GAC	CCC	192
	Gln	Pro	Val	Tyr	Ala	Arg	Asn	Ala	Asn	Phe	Gln	Arg	Ile	Thr	Asp	Pro	
		50					55				60						
25	AAC	GCC	AAG	TCC	GGC	TGC	GAT	GGC	GGC	TCC	GCC	TTC	TCC	TGC	GCC	GAC	240
	Asn	Ala	Lys	Ser	Gly	Cys	Asp	Gly	Gly	Ser	Ala	Phe	Ser	Cys	Ala	Asp	
	65					70					75				80		
30	CAG	ACC	CCG	TGG	GCC	GTG	AGC	GAC	GAC	TTT	GCC	TAC	GGT	TTC	GCG	GCT	288
	Gln	Thr	Pro	Trp	Ala	Val	Ser	Asp	Asp	Phe	Ala	Tyr	Gly	Phe	Ala	Ala	
					85					90					95		
35	ACG	GCG	CTC	GCC	GGC	CAG	TCC	GAG	TCT	TCG	TGG	TGC	TGT	GCC	TGC	TAC	336
	Thr	Ala	Leu	Ala	Gly	Gln	Ser	Glu	Ser	Ser	Trp	Cys	Cys	Ala	Cys	Tyr	
				100					105					110			
40	GAA	CTC	ACC	TTC	ACT	TCG	GGC	CCC	GTT	GCT	GGC	AAG	AAG	ATG	GCT	GTC	384
	Glu	Leu	Thr	Phe	Thr	Ser	Gly	Pro	Val	Ala	Gly	Lys	Lys	Met	Ala	Val	
				115				120					125				
45	CAG	TCC	ACC	AGC	ACT	GGC	GGT	GAC	CTC	GGT	AGC	AAC	CAC	TTT	GAC	CTC	432
	Gln	Ser	Thr	Ser	Thr	Gly	Gly	Asp	Leu	Gly	Ser	Asn	His	Phe	Asp	Leu	
		130				135						140					
50	AAC	ATG	CCA	GGT	GGC	GGT	GTC	GGC	ATC	TTC	GAC	GGC	TGC	TCG	CCT	CAG	480
	Asn	Met	Pro	Gly	Gly	Gly	Val	Gly	Ile	Phe	Asp	Gly	Cys	Ser	Pro	Gln	
	145					150					155				160		
55	GTT	GGC	GGT	CTC	GCC	GGC	CAG	CGC	TAT	GGC	GGC	GTC	TCG	TCC	CGC	AGC	528
	Val	Gly	Gly	Leu	Ala	Gly	Gln	Arg	Tyr	Gly	Gly	Val	Ser	Ser	Arg	Ser	
				165						170					175		
60	GAA	TGC	GAC	TCC	TTC	CCC	GCG	GCA	CTC	AAG	CCC	GGC	TGC	TAC	TGG	CGC	576
	Glu	Cys	Asp	Ser	Phe	Pro	Ala	Ala	Leu	Lys	Pro	Gly	Cys	Tyr	Trp	Arg	
				180					185					190			
65	TAC	GAC	TGG	TTT	AAG	AAC	GCC	GAC	AAT	CCG	AGC	TTC	AGC	TTC	CGT	CAG	624
	Tyr	Asp	Trp	Phe	Lys	Asn	Ala	Asp	Asn	Pro	Ser	Phe	Ser	Phe	Arg	Gln	
			195					200					205				
70	GTC	CAG	TGC	CCA	GCC	GAG	CTC	GTC	GCT	CGC	ACC	GGA	TGC	CGC	CGC	AAC	672
	Val	Gln	Cys	Pro	Ala	Glu	Leu	Val	Ala	Arg	Thr	Gly	Cys	Arg	Arg	Asn	
		210					215					220					
75	GAC	GAC	GGC	AAC	TTC	CCT	GCC	GTC	CAG	ATC	CCC	TCC	AGC	AGC	ACC	AGC	720
	Asp	Asp	Gly	Asn	Phe	Pro	Ala	Val	Gln	Ile	Pro	Ser	Ser	Ser	Thr	Ser	
	225					230					235				240		
80	TCT	CCG	GTC	AAC	CAG	CCT	ACC	AGC	ACC	AGC	ACC	ACG	TCC	ACC	TCC	ACC	768
	Ser	Pro	Val	Asn	Gln	Pro	Thr	Ser	Thr	Ser	Thr	Thr	Ser	Thr	Ser	Thr	
				245						250					255		

52

	ACC	TCG	AGC	CCG	CCA	GTC	CAG	CCT	ACG	ACT	CCC	AGC	GGC	TGC	ACT	GCT	816
	Thr	Ser	Ser	Pro	Pro	Val	Gln	Pro	Thr	Thr	Pro	Ser	Gly	Cys	Thr	Ala	
				260					265					270			
5	GAG	AGC	TGG	GCT	CAG	TGC	GGC	GGC	AAT	GGC	TGG	AGC	GGG	TGC	ACC	ACC	864
	Glu	Arg	Trp	Ala	Gln	Cys	Gly	Gly	Asn	Gly	Trp	Ser	Gly	Cys	Thr	Thr	
			275				280						285				
10	TGC	GTC	GCT	GGC	AGC	ACT	TGC	ACG	AAG	ATT	AAT	GAC	TGG	TAC	CAT	CAG	912
	Cys	Val	Ala	Gly	Ser	Thr	Cys	Thr	Lys	Ile	Asn	Asp	Trp	Tyr	His	Gln	
		290					295					300					
	TGC	CTG	TAG	A													
15	Cys	Leu	*														922
	305																
	(2) INFORMATION FOR SEQ ID NO: 66:																
20	(i) SEQUENCE CHARACTERISTICS:																
	(A) LENGTH: 307 amino acids																
	(B) TYPE: amino acid																
	(D) TOPOLOGY: linear																
	(ii) MOLECULE TYPE: protein																
25	(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:																
	Pro	Phe	Met	Met	Val	Ala	Trp	Trp	Ser	Leu	Phe	Leu	Tyr	Gly	Leu	Gln	
	1				5					10					15		
30	Val	Ala	Ala	Pro	Ala	Phe	Ala	Ala	Asp	Gly	Arg	Ser	Thr	Arg	Tyr	Trp	
				20					25					30			
	Asp	Cys	Cys	Lys	Pro	Ser	Cys	Ser	Trp	Pro	Gly	Lys	Ala	Leu	Val	Asn	
			35					40					45				
35	Gln	Pro	Val	Tyr	Ala	Arg	Asn	Ala	Asn	Phe	Gln	Arg	Ile	Thr	Asp	Pro	
	50						55					60					
	Asn	Ala	Lys	Ser	Gly	Cys	Asp	Gly	Gly	Ser	Ala	Phe	Ser	Cys	Ala	Asp	
40	65				70					75						80	
	Gln	Thr	Pro	Trp	Ala	Val	Ser	Asp	Asp	Phe	Ala	Tyr	Gly	Phe	Ala	Ala	
					85					90					95		
45	Thr	Ala	Leu	Ala	Gly	Gln	Ser	Glu	Ser	Ser	Trp	Cys	Cys	Ala	Cys	Tyr	
				100					105					110			
	Glu	Leu	Thr	Phe	Thr	Ser	Gly	Pro	Val	Ala	Gly	Lys	Lys	Met	Ala	Val	
			115					120					125				
50	Gln	Ser	Thr	Ser	Thr	Gly	Gly	Asp	Leu	Gly	Ser	Asn	His	Phe	Asp	Leu	
	130						135					140					
	Asn	Met	Pro	Gly	Gly	Gly	Val	Gly	Ile	Phe	Asp	Gly	Cys	Ser	Pro	Gln	
55	145				150					155						160	
	Val	Gly	Gly	Leu	Ala	Gly	Gln	Arg	Tyr	Gly	Gly	Val	Ser	Ser	Arg	Ser	
				165						170					175		
60	Glu	Cys	Asp	Ser	Phe	Pro	Ala	Ala	Leu	Lys	Pro	Gly	Cys	Tyr	Trp	Arg	
				180					185					190			
	Tyr	Asp	Trp	Phe	Lys	Asn	Ala	Asp	Asn	Pro	Ser	Phe	Ser	Phe	Arg	Gln	
			195					200					205				
65	Val	Gln	Cys	Pro	Ala	Glu	Leu	Val	Ala	Arg	Thr	Gly	Cys	Arg	Arg	Asn	
	210						215					220					
	Asp	Asp	Gly	Asn	Phe	Pro	Ala	Val	Gln	Ile	Pro	Ser	Ser	Ser	Thr	Ser	

SUBSTITUTE SHEET (RULE 26)

53

	225		230		235		240
	Ser Pro Val Asn Gln Pro Thr Ser Thr Ser Thr Thr Ser Thr Ser Thr						
		245			250		255
5	Thr Ser Ser Pro Pro Val Gln Pro Thr Thr Pro Ser Gly Cys Thr Ala						
		260		265			270
10	Glu Arg Trp Ala Gln Cys Gly Gly Asn Gly Trp Ser Gly Cys Thr Thr						
		275		280		285	
	Cys Val Ala Gly Ser Thr Cys Thr Lys Ile Asn Asp Trp Tyr His Gln						
		290		295		300	
15	Cys Leu *						
	305						
	(2) INFORMATION FOR SEQ ID NO: 68:						
	(i) SEQUENCE CHARACTERISTICS:						
20	(A) LENGTH: 922 base pairs						
	(B) TYPE: nucleic acid						
	(C) STRANDEDNESS: single						
	(D) TOPOLOGY: linear						
	(ii) MOLECULE TYPE: cDNA						
25	(ix) FEATURE:						
	(A) NAME/KEY: CDS						
	(B) LOCATION: 2..922						
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 67:						
30	C CCA TTT ATG ATG GTC GCG TGG TGG TCT CTA TTT CTG TAC GGC CTT						46
	Pro Phe Met Met Val Ala Trp Trp Ser Leu Phe Leu Tyr Gly Leu						
	1 5 10 15						
35	CAG GTC GCG GCA CCT GCT TTC GCT GCT GAT GGC AGG TCC ACG AGG TAC						94
	Gln Val Ala Ala Pro Ala Phe Ala Ala Asp Gly Arg Ser Thr Arg Tyr						
	20 25 30						
40	TGG GAT TGT TGT AAG CCC TCT TGC TCC TGG GGC GAC AAG GCC TCG GTC						142
	Trp Asp Cys Cys Lys Pro Ser Cys Ser Trp Gly Asp Lys Ala Ser Val						
	35 40 45						
45	AGC GCC CCC GTC CTG ACC TGC GAC AAG AAC GAC AAC CCC ATC TCC GAC						190
	Ser Ala Pro Val Leu Thr Cys Asp Lys Asn Asp Asn Pro Ile Ser Asp						
	50 55 60						
50	GCC AAC GCC GTG AGC GGT TGC AAC GGC GGC ACT TCC TAC ACC TGC AGC						238
	Ala Asn Ala Val Ser Gly Cys Asn Gly Gly Thr Ser Tyr Thr Cys Ser						
	65 70 75						
55	AAC AAC TCC CCG TGG GCT GTC AAC GAC AAC CTC GCC TAT GGC TTT GCC						286
	Asn Asn Ser Pro Trp Ala Val Asn Asp Asn Leu Ala Tyr Gly Phe Ala						
	80 85 90 95						
60	GCT ACC AAG CTC TCT GGA GGC TCC GAG TCC AGC TGG TGC TGT GCT TGC						334
	Ala Thr Lys Leu Ser Gly Gly Ser Glu Ser Ser Trp Cys Cys Ala Cys						
	100 105 110						
65	TAC GCT CTC ACC TTT ACG ACT GGC CCC GTG AAG GGC AAG ACC ATG GTC						382
	Tyr Ala Leu Thr Phe Thr Thr Gly Pro Val Lys Gly Lys Thr Met Val						
	115 120 125						
70	GTA CAG TCC ACC AAC ACC GGA GGC GAT CTC GGC GAG AAC CAC TTC GAT						430
	Val Gln Ser Thr Asn Thr Gly Gly Asp Leu Gly Glu Asn His Phe Asp						
	130 135 140						
75	CTC CAG ATG CCC GGC GGC GGT GTC GGC ATC TTT GAC GGC TGC AGC TCC						478
	Leu Gln Met Pro Gly Gly Gly Val Gly Ile Phe Asp Gly Cys Ser Ser						
	145 150 155						

SUBSTITUTE SHEET (RULE 26)

5	CAG Gln 160	TGG Trp	GGT Gly	GGC Gly	CTC Leu	GGC Gly	GGT Gly	GCT Ala	CAG Gln	TAC Tyr	GGC Gly	GGC Gly	ATC Ile	TCG Ser	TCG Ser	CGA Arg 175	526
10	AGC Ser	GAC Asp	TGC Cys	GAC Asp	AGC Ser	TTC Phe	CCC Pro	GAG Glu	CTG Leu	CTC Leu	AAG Lys	GAC Asp	GGC Gly	TGC Cys	TAC Tyr	TGG Trp 190	574
15	CGC Arg	TAC Tyr	GAC Asp	TGG Trp	TTC Phe	AAG Lys	AAC Asn	GCC Ala	GAC Asp	AAT Asn	CCG Pro	AGC Ser	TTC Phe	AGC Ser	TTC Phe	CGT Arg 205	622
20	CAG Gln	GTC Val	CAG Gln	TGC Cys	CCA Pro	GCC Ala	GAG Glu	CTC Leu	GTC Val	GCT Ala	CGC Arg	ACC Thr	GGA Gly	TGC Cys	CGC Arg	CGC Arg 220	670
25	AAC Asn	GAC Asp	GAC Asp	GGC Gly	AAC Asn	TTC Phe	CCT Pro	GCC Ala	GTC Val	CAG Gln	ATC Ile	CCC Pro	TCC Ser	AGC Ser	AGC Ser	ACC Thr 235	718
30	AGC Ser	TCT Ser	CCG Pro	GTC Val	AAC Asn	CAG Gln	CCT Pro	ACC Thr	AGC Ser	ACC Thr	AGC Thr	ACC Thr	ACG Thr	TCC Ser	ACC Thr	TCC Ser 255	766
35	ACC Thr	ACC Thr	TCG Ser	AGC Ser	CCG Pro	CCA Pro	GTC Val	CAG Gln	CCT Pro	ACG Thr	ACT Thr	CCC Pro	AGC Ser	GGC Gly	TGC Cys	ACT Thr 270	814
40	GCT Ala	GAG Glu	AGG Arg	TGG Trp	GCT Ala	CAG Gln	TGC Cys	GGC Gly	GGC Gly	AAT Asn	GGC Gly	TGG Trp	AGC Ser	GGC Gly	TGC Cys	ACC Thr 285	862
45	ACC Thr	TGC Cys	GTC Val	GCT Ala	GGC Gly	AGC Ser	ACT Thr	TGC Cys	ACG Thr	AAG Lys	ATT Ile	AAT Asn	GAC Asp	TGG Trp	TAC Tyr	CAT His 300	910
50	CAG Gln	TGC Cys	CTG Leu	TAG *													922
55	(2) INFORMATION FOR SEQ ID NO: 68:																
60	(i) SEQUENCE CHARACTERISTICS:																
65	(A) LENGTH: 307 amino acids																
70	(B) TYPE: amino acid																
75	(D) TOPOLOGY: linear																
80	(ii) MOLECULE TYPE: protein																
85	(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 68:																
90	Pro	Phe	Met	Met	Val	Ala	Trp	Trp	Ser	Leu	Phe	Leu	Tyr	Gly	Leu	Gln	
95	1				5					10					15		
100	Val	Ala	Ala	Pro	Ala	Phe	Ala	Ala	Asp	Gly	Arg	Ser	Thr	Arg	Tyr	Trp	
105				20					25					30			
110	Asp	Cys	Cys	Lys	Pro	Ser	Cys	Ser	Trp	Gly	Asp	Lys	Ala	Ser	Val	Ser	
115			35					40					45				
120	Ala	Pro	Val	Leu	Thr	Cys	Asp	Lys	Asn	Asp	Asn	Pro	Ile	Ser	Asp	Ala	
125			50				55					60					
130	Asn	Ala	Val	Ser	Gly	Cys	Asn	Gly	Gly	Thr	Ser	Tyr	Thr	Cys	Ser	Asn	
135	65				70					75						80	
140	Asn	Ser	Pro	Trp	Ala	Val	Asn	Asp	Asn	Leu	Ala	Tyr	Gly	Phe	Ala	Ala	
145				85					90						95		
150	Thr	Lys	Leu	Ser	Gly	Gly	Ser	Glu	Ser	Ser	Trp	Cys	Cys	Ala	Cys	Tyr	

55

	100	105	110
	Ala Leu Thr Phe Thr Thr Gly Pro Val Lys Gly Lys Thr Met Val Val		
	115	120	125
5	Gln Ser Thr Asn Thr Gly Gly Asp Leu Gly Glu Asn His Phe Asp Leu		
	130	135	140
10	Gln Met Pro Gly Gly Gly Val Gly Ile Phe Asp Gly Cys Ser Ser Gln		
	145	150	155
	Trp Gly Gly Leu Gly Gly Ala Gln Tyr Gly Gly Ile Ser Ser Arg Ser		
		165	170
15	Asp Cys Asp Ser Phe Pro Glu Leu Leu Lys Asp Gly Cys Tyr Trp Arg		
		180	185
	Tyr Asp Trp Phe Lys Asn Ala Asp Asn Pro Ser Phe Ser Phe Arg Gln		
	195	200	205
20	Val Gln Cys Pro Ala Glu Leu Val Ala Arg Thr Gly Cys Arg Arg Asn		
	210	215	220
25	Asp Asp Gly Asn Phe Pro Ala Val Gln Ile Pro Ser Ser Ser Thr Ser		
	225	230	235
	Ser Pro Val Asn Gln Pro Thr Ser Thr Ser Thr Thr Ser Thr Ser Thr		
		245	250
30	Thr Ser Ser Pro Pro Val Gln Pro Thr Thr Pro Ser Gly Cys Thr Ala		
		260	265
	Glu Arg Trp Ala Gln Cys Gly Gly Asn Gly Trp Ser Gly Cys Thr Thr		
	275	280	285
35	Cys Val Ala Gly Ser Thr Cys Thr Lys Ile Asn Asp Trp Tyr His Gln		
	290	295	300
	Cys Leu *		
40	305		

- (2) INFORMATION FOR SEQ ID NO: 69:
- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 928 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: cDNA
- (ix) FEATURE:
- (A) NAME/KEY: CDS
- (B) LOCATION: 1..928
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 69:

55	CCA TTT ATG ATG GTC GCG TGG TGG TCT CTA TTT CTG TAC GGC CTT CAG	48
	Pro Phe Met Met Val Ala Trp Trp Ser Leu Phe Leu Tyr Gly Leu Gln	
	1 5 10 15	
60	GTC GCG GCA CCT GCT TTC GCT GCT GAT GGC AGG TCC ACG AGG TAC TGG	96
	Val Ala Ala Pro Ala Phe Ala Ala Asp Gly Arg Ser Thr Arg Tyr Trp	
	20 25 30	
65	GAT TGC TGC AAG CCC TCT TGC TCT TGG GGC GGA AAG GCT GCT GTC AGC	144
	Asp Cys Cys Lys Pro Ser Cys Ser Trp Gly Gly Lys Ala Ala Val Ser	
	35 40 45	
	GCC CCT GCT TTG ACC TGT GAC AAG AAG GAC AAC CCC ATC TCA AAC CTG	192
	Ala Pro Ala Leu Thr Cys Asp Lys Lys Asp Asn Pro Ile Ser Asn Leu	
	50 55 60	

SUBSTITUTE SHEET (RULE 26)

[illegible]

(2) INFORMATION FOR SEQ ID NO: 70:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 309 amino acids
 (B) TYPE: amino acid

(D) TOPOLOGY: linear																	
(ii) MOLECULE TYPE: protein																	
(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 70:																	
5	Pro	Phe	Met	Met	Val	Ala	Trp	Trp	Ser	Leu	Phe	Leu	Tyr	Gly	Leu	Gln	
	1				5					10					15		
	Val	Ala	Ala	Pro	Ala	Phe	Ala	Ala	Asp	Gly	Arg	Ser	Thr	Arg	Tyr	Trp	
				20					25					30			
10	Asp	Cys	Cys	Lys	Pro	Ser	Cys	Ser	Trp	Gly	Gly	Lys	Ala	Ala	Val	Ser	
			35					40					45				
	Ala	Pro	Ala	Leu	Thr	Cys	Asp	Lys	Lys	Asp	Asn	Pro	Ile	Ser	Asn	Leu	
		50					55					60					
15	Asn	Ala	Val	Asn	Gly	Cys	Glu	Gly	Gly	Gly	Ser	Ala	Phe	Ala	Cys	Thr	
	65					70					75					80	
20	Asn	Tyr	Ser	Pro	Trp	Ala	Val	Asn	Asp	Asn	Leu	Ala	Tyr	Gly	Phe	Ala	
					85					90					95		
	Ala	Thr	Lys	Leu	Ala	Gly	Gly	Ser	Glu	Gly	Ser	Trp	Cys	Cys	Ala	Cys	
				100					105					110			
25	Tyr	Ala	Leu	Thr	Phe	Thr	Thr	Gly	Pro	Val	Lys	Gly	Lys	Thr	Met	Val	
			115					120					125				
	Val	Gln	Ser	Thr	Asn	Thr	Gly	Gly	Asp	Leu	Gly	Asp	Asn	His	Phe	Asp	
		130					135					140					
30	Leu	Met	Met	Pro	Gly	Gly	Gly	Val	Gly	Ile	Phe	Asp	Gly	Cys	Thr	Ser	
	145					150					155					160	
35	Gln	Phe	Gly	Lys	Ala	Leu	Gly	Gly	Ala	Gln	Tyr	Gly	Gly	Ile	Ser	Ser	
					165					170					175		
	Arg	Ser	Glu	Cys	Asp	Ser	Phe	Pro	Glu	Thr	Leu	Lys	Asp	Gly	Cys	His	
				180					185					190			
40	Trp	Arg	Phe	Asp	Trp	Phe	Lys	Asn	Ala	Asp	Asn	Pro	Ser	Phe	Ser	Phe	
			195					200					205				
	Arg	Gln	Val	Gln	Cys	Pro	Ala	Glu	Leu	Val	Ala	Arg	Thr	Gly	Cys	Arg	
		210					215					220					
45	Arg	Asn	Asp	Asp	Gly	Asn	Phe	Pro	Ala	Val	Gln	Ile	Pro	Ser	Ser	Ser	
	225					230					235					240	
50	Thr	Ser	Ser	Pro	Val	Asn	Gln	Pro	Thr	Ser	Thr	Ser	Thr	Thr	Ser	Thr	
					245					250					255		
	Ser	Thr	Thr	Ser	Ser	Pro	Pro	Val	Gln	Pro	Thr	Thr					

65 (2) INFORMATION FOR SEQ ID NO: 71:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 915 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

58

(ii) MOLECULE TYPE: cDNA

(ix) FEATURE:

(A) NAME/KEY: CDS

(B) LOCATION:1..915

5	(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 71:																
	ATG	ATG	GTC	CGC	TGG	TGG	TCT	CTA	TTT	CTG	TAC	GGC	CTT	CAG	GTC	GCG	48
	Met	Met	Val	Ala	Trp	Trp	Ser	Leu	Phe	Leu	Tyr	Gly	Leu	Gln	Val	Ala	
10	1				5					10					15		
	GCA	CCT	GCT	TTC	GCT	GCT	GAT	GGC	AGG	TCC	ACG	AGG	TAT	TGG	GAT	TGT	96
	Ala	Pro	Ala	Phe	Ala	Ala	Asp	Gly	Arg	Ser	Thr	Arg	Tyr	Trp	Asp	Cys	
				20					25					30			
15	TGC	AAG	CCG	TCA	TGT	GCT	TGG	TCC	GGC	AAG	GCC	TCA	GTG	TCA	TCT	CCC	144
	Cys	Lys	Pro	Ser	Cys	Ala	Trp	Ser	Gly	Lys	Ala	Ser	Val	Ser	Ser	Pro	
			35					40					45				
20	GTG	CGA	ACC	TGT	GAC	GCA	AAC	AAC	TCG	CCG	CTG	TCC	GAC	GTC	GAC	GCA	192
	Val	Arg	Thr	Cys	Asp	Ala	Asn	Asn	Ser	Pro	Leu	Ser	Asp	Val	Asp	Ala	
		50					55					60					
25	AAG	AGT	GCG	TGC	GAT	GGA	GGC	GTT	GCT	TAC	ACT	TGT	TCA	AAC	AAC	GCG	240
	Lys	Ser	Ala	Cys	Asp	Gly	Gly	Val	Ala	Tyr	Thr	Cys	Ser	Asn	Asn	Ala	
	65					70				75					80		
30	CCT	TGG	GCT	GTT	AAC	GAT	AAC	CTC	TCT	TAT	GGT	TTC	GCG	GCC	ACA	GCT	288
	Pro	Trp	Ala	Val	Asn	Asp	Asn	Leu	Ser	Tyr	Gly	Phe	Ala	Ala	Thr	Ala	
				85						90					95		
	ATC	AAT	GGC	GGC	AGC	GAG	TCT	AGC	TGG	TGC	TGT	GCA	TGC	TAC	AAG	TTG	336
	Ile	Asn	Gly	Gly	Ser	Glu	Ser	Ser	Trp	Cys	Cys	Ala	Cys	Tyr	Lys	Leu	
				100					105					110			
35	ACT	TTC	ACG	AGC	GGA	CCT	GCT	TCT	GGA	AAG	GTC	ATC	GTC	GTT	CAA	TCA	384
	Thr	Phe	Thr	Ser	Gly	Pro	Ala	Ser	Gly	Lys	Val	Met	Val	Val	Gln	Ser	
			115					120					125				
40	ACC	AAC	ACC	GGG	TAC	GAT	CTC	TCT	AAC	AAC	CAC	TTT	GAC	ATT	CTT	ATG	432
	Thr	Asn	Thr	Gly	Tyr	Asp	Leu	Ser	Asn	Asn	His	Phe	Asp	Ile	Leu	Met	
			130				135					140					
45	CCA	GGT	GGC	GGT	GTT	GGA	GCG	TTC	GAC	GGC	TGC	TCT	AGG	CAG	TAC	GGC	480
	Pro	Gly	Gly	Gly	Val	Gly	Ala	Phe	Asp	Gly	Cys	Ser	Arg	Gln	Tyr	Gly	
						150					155					160	
50	AGC	ATC	CCT	GGG	GAG	CGA	TAT	GGG	GGT	GTC	ACA	TCA	AGG	GAC	CAA	TGC	528
	Ser	Ile	Pro	Gly	Glu	Arg	Tyr	Gly	Gly	Val	Thr	Ser	Arg	Asp	Gln	Cys	
				165						170					175		
	GAC	CAA	ATG	CCA	AGT	GCA	CTC	AAG	CAG	GGC	TGC	TAT	TGG	CGC	TTC	GAT	576
	Asp	Gln	Met	Pro	Ser	Ala	Leu	Lys	Gln	Gly	Cys	Tyr	Trp	Arg	Phe	Asp	
				180				185						190			
55	TGG	TTC	AAG	AAC	GCC	GAC	AAT	CCG	AGC	TTC	AGC	TTC	CGT	CAG	GTC	CAG	624
	Trp	Phe	Lys	Asn	Ala	Asp	Asn	Pro	Ser	Phe	Ser	Phe	Arg	Gln	Val	Gln	
			195					200					205				
60	TGC	CCA	GCC	GAG	CTC	GTC	GCT	CGC	ACC	GGA	TGC	CGC	CGC	AAC	GAC	GAC	672
	Cys	Pro	Ala	Glu	Leu	Val	Ala	Arg	Thr	Gly	Cys	Arg	Arg	Asn	Asp	Asp	
			210				215					220					
65	GGC	AAC	TTC	CCT	GCC	GTC	CAG	ATC	CCC	TCC	AGC	AGC	ACC	AGC	TCT	CCG	720
	Gly	Asn	Phe	Pro	Ala	Val	Gln	Ile	Pro	Ser	Ser	Ser	Thr	Ser	Ser	Pro	
						230					235					240	
	GTC	AAC	CAG	CCT	ACC	AGC	ACC	AGC	ACC	ACG	TCC	ACC	TCC	ACC	ACC	TCG	768
	Val	Asn	Gln	Pro	Thr	Ser	Thr	Ser	Thr	Thr	Ser	Thr	Ser	Thr	Thr	Ser	
						245				250						255	

SUBSTITUTE SHEET (RULE 26)

59

	AGC CCG CCA GTC CAG CCT ACG ACT CCC AGC GGC TGC ACT GCT GAG AGG	816
	Ser Pro Pro Val Gln Pro Thr Thr Pro Ser Gly Cys Thr Ala Glu Arg	
	260 265 270	
5	TGG GCT CAG TGC GGC GGC AAT GGC TGG AGC GGC TGC ACC ACC TGC GTC	864
	Trp Ala Gln Cys Gly Gly Asn Gly Trp Ser Gly Cys Thr Thr Cys Val	
	275 280 285	
10	GCT GGC AGC ACT TGC ACG AAG ATT AAT GAC TGG TAC CAT CAG TGC CTC	912
	Ala Gly Ser Thr Cys Thr Lys Ile Asn Asp Trp Tyr His Gln Cys Leu	
	290 295 300	
	TAG	
15	*	915
	305	
	(2) INFORMATION FOR SEQ ID NO: 72:	
20	(i) SEQUENCE CHARACTERISTICS:	
	(A) LENGTH: 305 amino acids	
	(B) TYPE: amino acid	
	(D) TOPOLOGY: linear	
	(ii) MOLECULE TYPE: protein	
25	(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 72:	
	Met Met Val Ala Trp Trp Ser Leu Phe Leu Tyr Gly Leu Gln Val Ala	
	1 5 10 15	
30	Ala Pro Ala Phe Ala Ala Asp Gly Arg Ser Thr Arg Tyr Trp Asp Cys	
	20 25 30	
	Cys Lys Pro Ser Cys Ala Trp Ser Gly Lys Ala Ser Val Ser Ser Pro	
	35 40 45	
35	Val Arg Thr Cys Asp Ala Asn Asn Ser Pro Leu Ser Asp Val Asp Ala	
	50 55 60	
40	Lys Ser Ala Cys Asp Gly Gly Val Ala Tyr Thr Cys Ser Asn Asn Ala	
	65 70 75 80	
	Pro Trp Ala Val Asn Asp Asn Leu Ser Tyr Gly Phe Ala Ala Thr Ala	
	85 90 95	
45	Ile Asn Gly Gly Ser Glu Ser Ser Trp Cys Cys Ala Cys Tyr Lys Leu	
	100 105 110	
	Thr Phe Thr Ser Gly Pro Ala Ser Gly Lys Val Met Val Val Gln Ser	
	115 120 125	
50	Thr Asn Thr Gly Tyr Asp Leu Ser Asn Asn His Phe Asp Ile Leu Met	
	130 135 140	
	Pro Gly Gly Gly Val Gly Ala Phe Asp Gly Cys Ser Arg Gln Tyr Gly	
	145 150 155 160	
55	Ser Ile Pro Gly Glu Arg Tyr Gly Gly Val Thr Ser Arg Asp Gln Cys	
	165 170 175	
60	Asp Gln Met Pro Ser Ala Leu Lys Gln Gly Cys Tyr Trp Arg Phe Asp	
	180 185 190	
	Trp Phe Lys Asn Ala Asp Asn Pro Ser Phe Ser Phe Arg Gln Val Gln	
	195 200 205	
65	Cys Pro Ala Glu Leu Val Ala Arg Thr Gly Cys Arg Arg Asn Asp Asp	
	210 215 220	
	ly Asn Phe Pro Ala Val Gln Ile Pro Ser Ser Ser Thr Ser Ser Pro	

SUBSTITUTE SHEET (RULE 26)

60

	225		230		235		240
	Val Asn Gln Pro Thr Ser Thr Ser Thr Thr Ser Thr Ser Thr Thr Ser						
			245		250		255
5	Ser Pro Pro Val Gln Pro Thr Thr Pro Ser Gly Cys Thr Ala Glu Arg						
			260		265		270
10	Trp Ala Gln Cys Gly Gly Asn Gly Trp Ser Gly Cys Thr Thr Cys Val						
			275		280		285
	Ala Gly Ser Thr Cys Thr Lys Ile Asn Asp Trp Tyr His Gln Cys Leu						
			290		295		300
15	*						
	305						
	(2) INFORMATION FOR SEQ ID NO: 73:						
20	(i) SEQUENCE CHARACTERISTICS:						
	(A) LENGTH: 925 base pairs						
	(B) TYPE: nucleic acid						
	(C) STRANDEDNESS: single						
	(D) TOPOLOGY: linear						
25	(ii) MOLECULE TYPE: cDNA						
	(ix) FEATURE:						
	(A) NAME/KEY: CDS						
	(B) LOCATION: 2..925						
	(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 73:						
30	C CCA TTT ATG ATG GTC GCG TGG TGG TCT CTA TTT CTG TAC GGC CTT						46
	Pro Phe Met Met Val Ala Trp Trp Ser Leu Phe Leu Tyr Gly Leu						
	1 5 10 15						
35	CAG GTC GCG GCA CCT GCT TTC GCT GCT GAT GGC AGG TCC ACG CGG TAT						94
	Gln Val Ala Ala Pro Ala Phe Ala Ala Asp Gly Arg Ser Thr Arg Tyr						
	20 25 30						
40	TGG GAT TGC TGT AAG CCC AGC TGC TCC TGG CCC GAC AAG GCC CCC GTA						142
	Trp Asp Cys Cys Lys Pro Ser Cys Ser Trp Pro Asp Lys Ala Pro Val						
	35 40 45						
45	GGT TCC CCC GTA GGC ACC TGC GAC GCC GGC AAC AGC CCC CTC GGC GAC						190
	Gly Ser Pro Val Gly Thr Cys Asp Ala Gly Asn Ser Pro Leu Gly Asp						
	50 55 60						
50	CCC CTG GCC AAG TCT GGC TGC GAG GGC GGC CCG TCG TAC ACG TGC GCC						238
	Pro Leu Ala Lys Ser Gly Cys Glu Gly Gly Pro Ser Tyr Thr Cys Ala						
	65 70 75						
55	AAC TAC CAG CCG TGG GCG GTC AAC GAC CAG CTG GCC TAC GGC TTC GCG						286
	Asn Tyr Gln Pro Trp Ala Val Asn Asp Gln Leu Ala Tyr Gly Phe Ala						
	80 85 90 95						
60	GCC ACG GCC ATC AAC GGC GGC ACC GAG GAC TCG TGG TGC TGC GCC TGC						334
	Ala Thr Ala Ile Asn Gly Gly Thr Glu Asp Ser Trp Cys Cys Ala Cys						
	100 105 110						
65	TAC AAG CTC ACC TTC ACC GAC GGC CCG GCC TCG GGC AAG ACC ATG ATC						382
	Tyr Lys Leu Thr Phe Thr Asp Gly Pro Ala Ser Gly Lys Thr Met Ile						
	115 120 125						
70	GTC CAG TCC ACC AAC ACG GGC GGC GAC CTG TCC GAC AAC CAC TTC GAC						430
	Val Gln Ser Thr Asn Thr Gly Gly Asp Leu Ser Asp Asn His Phe Asp						
	130 135 140						
75	CTG CTC ATC CCC GGC GGC GGC GTC GGC ATC TTC GAC GGC TGC ACC TCC						478
	Leu Leu Ile Pro Gly Gly Gly Val Gly Ile Phe Asp Gly Cys Thr Ser						
	145 150 155						

SUBSTITUTE SHEET (RULE 26)

61

5	CAG 160	TAC Tyr	GGC Gly	CAG Gln	GCC Ala	CTG Leu	CCC Pro	GGC Gly	GCC Ala	CAG Gln	TAC Tyr	GGC Gly	GGC Gly	GTC Val	AGC Ser	TCC Ser	526
10	CGC Arg	GCC Ala	GAG Glu	TGC Cys	GAC Asp	CAG Gln	ATG Met	CCC Pro	GAG Glu	GCC Ala	ATC Ile	AAG Lys	GCC Ala	GGC Gly	TGC Cys	CAG Gln	574
15	TGG Trp	CGC Arg	TAC Tyr	GAT Asp	TGG Trp	TTT Phe	AAG Lys	AAC Asn	GCC Ala	GAC Asp	AAT Asn	CCG Pro	AGC Ser	TTC Phe	AGC Ser	TTC Phe	622
20	CGT Arg	CAG Gln	GTC Val	CAG Gln	TGC Cys	CCA Pro	GCC Ala	GAG Glu	CTC Leu	GTC Val	GCT Ala	CGC Arg	ACC Thr	GGA Gly	TGC Cys	CGC Arg	670
25	CGC Arg	AAC Asn	GAC Asp	GAC Asp	GGC Gly	AAC Asn	TTC Phe	CCT Pro	GCC Ala	GTC Val	CAG Gln	ATC Ile	CCC Pro	TCC Ser	AGC Ser	AGC Ser	718
30	ACC Thr	AGC Ser	TCT Ser	CCG Pro	GTC Val	AAC Asn	CAG Gln	CCT Pro	ACC Thr	AGC Ser	ACC Thr	AGC Ser	ACC Thr	ACG Thr	TCC Ser	ACC Thr	766
35	TCC Ser	ACC Thr	ACC Thr	TCG Ser	AGC Ser	CCG Pro	CCA Pro	GTC Val	CAG Gln	CCT Pro	ACG Thr	ACT Thr	CCC Pro	AGC Ser	GGC Gly	TGC Cys	814
40	ACT Thr	GCT Ala	GAG Glu	AGG Arg	TGG Trp	GCT Ala	CAG Gln	TGC Cys	GGC Gly	GGC Gly	AAT Asn	GGC Gly	TGG Trp	AGC Ser	GGC Gly	TGC Cys	862
45	ACC Thr	ACC Thr	TGC Cys	GTC Val	GCT Ala	GGC Gly	AGC Ser	ACT Thr	TGC Cys	ACG Thr	AAG Lys	ATT Ile	AAT Asn	GAC Asp	TGG Trp	TAC Tyr	910
50	CAT His	CAG Gln	TGC Cys	CTG Leu	TAG *												925

(2) INFORMATION FOR SEQ ID NO: 74:

45	(i) SEQUENCE CHARACTERISTICS:
	(A) LENGTH: 308 amino acids
	(B) TYPE: amino acid
	(D) TOPOLOGY: linear
	(ii) MOLECULE TYPE: protein
50	(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 74:
55	Pro Phe Met Met Val Ala Trp Trp Ser Leu Phe Leu Tyr Gly Leu Gln 1 5 10 15
60	Val Ala Ala Pro Ala Phe Ala Ala Asp Gly Arg Ser Thr Arg Tyr Trp 20 25 30
65	Asp Cys Cys Lys Pro Ser Cys Ser Trp Pro Asp Lys Ala Pro Val Gly 35 40 45
70	Ser Pro Val Gly Thr Cys Asp Ala Gly Asn Ser Pro Leu Gly Asp Pro 50 55 60
75	Leu Ala Lys Ser Gly Cys Glu Gly Gly Pro Ser Tyr Thr Cys Ala Asn 65 70 75 80
80	Tyr Gln Pro Trp Ala Val Asn Asp Gln Leu Ala Tyr Gly Phe Ala Ala 85 90 95
85	Thr Ala Ile Asn Gly Gly Thr Glu Asp Ser Trp Cys Cys Ala Cys Tyr

SUBSTITUTE SHEET (RULE 26)

62

	100	105	110
	Lys Leu Thr Phe Thr Asp Gly	Pro Ala Ser Gly	Lys Thr Met Ile Val
	115	120	125
5	Gln Ser Thr Asn Thr Gly Gly	Asp Leu Ser Asp	Asn His Phe Asp Leu
	130	135	140
10	Leu Ile Pro Gly Gly Gly Val	Gly Ile Phe Asp	Gly Cys Thr Ser Gln
	145	150	155
	Tyr Gly Gln Ala Leu Pro Gly	Ala Gln Tyr Gly	Gly Val Ser Ser Arg
	165	170	175
15	Ala Glu Cys Asp Gln Met Pro	Glu Ala Ile Lys Ala	Gly Cys Gln Trp
	180	185	190
	Arg Tyr Asp Trp Phe Lys Asn	Ala Asp Asn Pro Ser	Phe Ser Phe Arg
	195	200	205
20	Gln Val Gln Cys Pro Ala Glu	Leu Val Ala Arg	Thr Gly Cys Arg Arg
	210	215	220
25	Asn Asp Asp Gly Asn Phe Pro	Ala Val Gln Ile	Pro Ser Ser Ser Thr
	225	230	235
	Ser Ser Pro Val Asn Gln Pro	Thr Ser Thr Ser	Thr Ser Thr Ser
	245	250	255
30	Thr Thr Ser Ser Pro Pro Val	Gln Pro Thr Thr	Pro Ser Gly Cys Thr
	260	265	270
	Ala Glu Arg Trp Ala Gln Cys	Gly Gly Asn Gly	Trp Ser Gly Cys Thr
	275	280	285
35	Thr Cys Val Ala Gly Ser Thr	Cys Thr Lys Ile	Asn Asp Trp Tyr His
	290	295	300
40	Gln Cys Leu *		
	305		

PATENT CLAIMS

1. A method for providing a novel DNA sequence encoding a polypeptide from a micro-organism with an activity of interest
5 comprises the following steps:
i) PCR amplification of said DNA with PCR primers with homology to (a) known gene(s) encoding a polypeptide with an activity of interest,
ii) linking the obtained PCR product to a 5' structural gene
10 sequence and a 3' structural gene sequence,
iii) expressing said resulting hybrid DNA sequence,
iv) screening for hybrid DNA sequences encoding a polypeptide with said activity of interest or related activity,
v) isolating the hybrid DNA sequence identified in step iv)
15
2. The method according to claim 1 wherein the PCR primers in step i) have homology to conserved regions in (a) known structural gene(s) or the polypeptide(s) thereof.
- 20 3. The method according to claim 1 wherein the PCR primers in step i) are degenerated on the basis of conserved regions in (a) known gene(s).
4. The method according to any of claims 1 to 3 wherein the PCR
25 amplification in step i) is performed using naturally occurring DNA as template.
5. The method according to any of claims 1 to 3 wherein the microorganism has not been subjected to "in vitro" selection.
30
6. The method according to any of claims 1 to 5 wherein the PCR amplification in step i) is performed on a sample containing DNA from an un-isolated microorganism.
- 35 7. The method according to any of claims 1 to 6 wherein the 5' and 3' structural gene sequences originate from two different structural genes encoding polypeptides having the same activity.

8. The method according to any of claims 1 to 7 wherein the 5' structural gene sequence and the 3' structural gene sequence originate from the same structural gene sequence.

5

9. The method according to any of claims 1 to 8 wherein the 5' structural gene sequence and the 3' structural gene sequence originate from two different structural gene sequences encoding polypeptides having different activities.

10

10. The method according to any of claims 1 to 9 comprising the following steps:

- i) PCR amplification of DNA from micro-organisms with PCR primers being homologous to conserved regions of
15 a known gene encoding a polypeptide with an activity of interest,
- ii) cloning the obtained PCR product into a gene encoding a polypeptide having the activity of interest, where said gene is not identical to the gene from which the PCR
20 product is obtained, which gene is situated in an expression vector,
- iii) transforming said expression vector into a suitable host cell,
- iiia) culturing said host cell under suitable conditions,
- 25 iv) screening for clones comprising a DNA sequence originated from the PCR amplification in step i) encoding a polypeptide with said activity of interest or related activity,
- v) isolating the DNA sequence identified in step iv).

30

11. The method according to claims 1 to 10, wherein the micro-organism from which DNA is to be PCR amplified in step i) is a prokaryote or an eukaryote.

35 12. The method according to any of claims 1 to 11, wherein the PCR amplification in step i) is performed on DNA from an uncultivable organism.

13. The method according to claim 12, wherein the un-cultivable organism is an algae, a fungi or a protozoa.
- 5 14. The method according to claims 12 and 13, wherein said un-cultivable organism is from the group of extremophiles and planktonic marine organisms.
15. The method according to any of claims 1 to 11, wherein the
10 PCR amplification in step i) is performed on DNA from a cultivable organism.
16. The method according to claim 15, wherein said cultivable organism is selected from the group of bacteria, fungal
15 organisms, such as filamentous fungi or yeasts.
17. The method according to claim 16, wherein said PCR amplification in step i) is performed on one or more polynucleotides comprised in a vector, plasmid or the like, such as on a cDNA
20 library from cultivable organisms.
18. The method according any of claims 1 to 17, wherein said activity of interest is an enzymatic activity.
- 25 19. The method according to claim 18, wherein said enzyme activity is selected from the group comprising phosphatases oxidoreductases, transferases, hydrolases, such as esterases, in particular lipases and phytases, such as glucosidases, in particular xylanases, cellulases, hemicellulases, and amylases,
30 such as peptidases, in particular proteases, lyases, isomerases and ligases.
20. The method according to any of claims 10 to 19, wherein said host cell mentioned under iii) of claim 10 is a micro-organism,
35 preferably a yeast or a bacteria.
21. The method according to claim 20, wherein said host cell is a yeast such as a strain of *Saccharomyces*, in particular

Saccharomyces cerevisiae.

22. The method according to claim 20, wherein said host cell is a bacteria such as a strain of *Bacillus*, in particular of
5 *Bacillus subtilis*, or a strain *Escherichia coli*.

23. The method according to any of claims 1 to 22, wherein the clones/hybrid DNA sequences mentioned in step iv), are screened for enzymatic activity.

10

24. The method according to claim 23, wherein the screened clones/hybrid DNA sequences are tested for wash performance.

25. A novel DNA sequence provided according to any of the method
15 claims 1 to 24.

26. A polypeptide with an activity of interest encoded by a DNA sequence of claim 25.

1/4

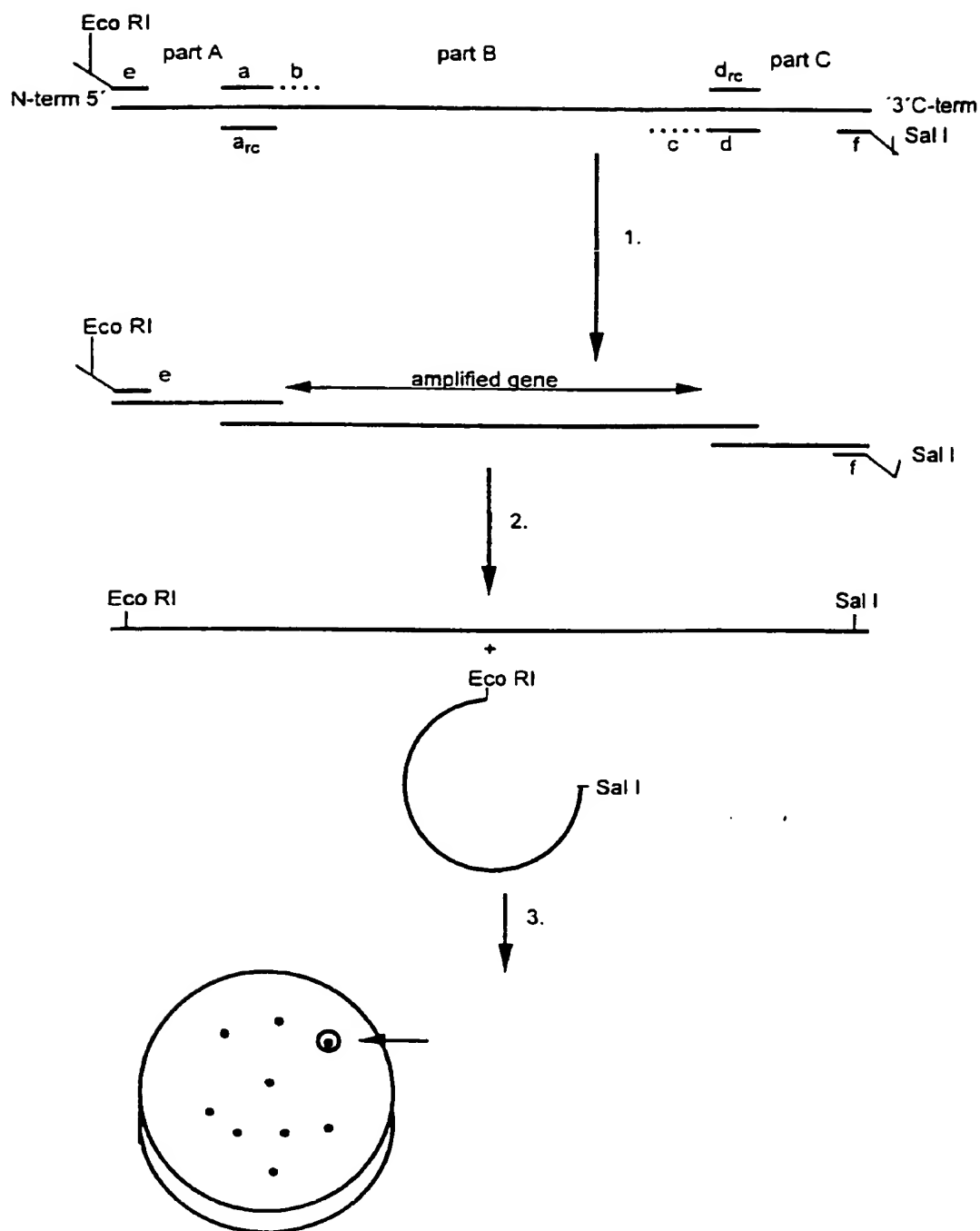


Figure 1

SUBSTITUTE SHEET (RULE 26)

PULPYME_L	1	-----MRQK-----	KLTFILAFVCF	17
XYNA_BACCI	1	-----MFKFKKN-----	-----FLV	10
XYNA_BACPU	1	-----MNLRLK-----	RLLFVMCIGLTL	19
XYNA_BACST	1	-----MKLKKK-----	-----MLT	9
XYNA_BACSU	1	-----MFKFKKN-----	-----FLV	10
XYNA_CLOAB	1	-----MLRRK-----	VIFTVLATLVMTS	18
XYNA_CLOSR	1	-----MKRKVKKM-----	AAMATS I IMAIMI	21
XYNB_STRLI	1	---MNL LVQPRRRR---GPV TLLVR---	SAWAVALAALALM	34
XYNC_STRLI	1	MQQDGTQQDR IKQSPAPLNGMSRRGFLGGAGTLALATASGLL	42	
PULPYME_L	18	LTLP AE-----	IIQAQ	28
XYNA_BACCI	11	GLSAAL-----	MSI	19
XYNA_BACPU	20	LTAVP-----	AHAR	28
XYNA_BACST	10	LLLTAS-----	MSF	18
XYNA_BACSU	11	GLSAAL-----	MSI	19
XYNA_CLOAB	19	LTIVDNTAFAATNLNTTESTFSKEVLSTQKTYSAFNTQAAPK	60	
XYNA_CLOSR	22	ILHSIP-----	VLAGR	32
XYNB_STRLI	35	LPGT AQ-----	ADT	43
XYNC_STRLI	43	LPGT AH-----	AAT	51
PULPYME_L	29	IVTOMSGNHGQYDYEFWKDSGGSGTILNHSGTFSACNNV	70	
XYNA_BACCI	20	SLFSATASAAS TDYWQNWTDGGGIVNAVNGSGGNYGVNSIT	61	
XYNA_BACPU	29	TITNMEMNHSGYDYELWKDYGLNTSITLNNCGAFSAGNNI	69	
XYNA_BACST	19	GLFGATSSAA-TDYWQYATDGGGMVNAVNGPSSNYGVNTQNT	59	
XYNA_BACSU	20	SLFSATASAAS TDYWQNWTDGGGIVNAVNGSGGNYGVNSIT	61	
XYNA_CLOAB	61	TITSNIEIGVNGGYDYELWKDYGLNTSITLKNCGAFSCQNSI	101	
XYNA_CLOSR	33	IIYDNEGTTHGQYDYELWKDYGLNTIIVELNDGGETFGCQNSI	73	
XYNB_STRLI	44	VVTTNQEGTNMGYYYSFTDSQGTVSNNMGSGLQYQTSVRI	85	
XYNC_STRLI	52	TITTNQGTGT-DEMYMSFTDGGGSSVSTLNGGSSYSTQNTNC	92	
PULPYME_L	71	MLILFRKGGKKFNETQTHQQVGNMSILYGANFQ-DENERAEE	CV 111	
XYNA_BACCI	62	GNFVVGKQWTTGS-----PFRITINAGVWAPNGGKGTTL	96	
XYNA_BACPU	70	GNALFRKGGKKFDSTRTHQLGNISINYNASFN-EGGNSITCV	110	
XYNA_BACST	60	GNFVVGKQWTVGS-----PNRVINAGIWEBSCGKGTTL	94	
XYNA_BACSU	62	GNFVVGKQWTTGS-----PFRITINAGVWAPNGGKGTTL	96	
XYNA_CLOAB	102	GNALFRKGGKKFNDTQTYKQLGNISVYDCNYQ-EGYNSITCV	142	
XYNA_CLOSR	74	GNALFRKGGKKFNSDKTYQELGDIVVEYGCDDYN-ENGSITCV	114	
XYNB_STRLI	86	GNFVAGKSWANG-----GRRTVCSGSFN-EGSITCV	118	
XYNC_STRLI	93	GNFVAGKSWANG-----GN-VRNGYFN-EGGCL	124	
PULPYME_L	112	YGVVVDPEVEYVYVSEWGNWREPEATPKGTITVDEG	152	
XYNA_BACCI	97	YGVVTRSEIIEYVYVDSWGTTEETGTYKG-IVKSSBEG	136	
XYNA_BACPU	111	YGVVQSSELAIEYVYVDSWGTTEETGTYKG-IVKSSBEG	150	
XYNA_BACST	95	YGVVTRSEIIEYVYVDSWGTTEETGTYKG-IVKSSBEG	135	
XYNA_BACSU	97	YGVVTRSEIIEYVYVDSWGTTEETGTYKG-IVKSSBEG	136	
XYNA_CLOAB	143	YGVVTRSEIIEYVYVDSWGTTEETGTYKG-IVKSSBEG	183	
XYNA_CLOSR	115	YGVVTRSEIIEYVYVDSWGTTEETGTYKG-IVKSSBEG	156	
XYNB_STRLI	119	YGVVTRSEIIEYVYVDSWGTTEETGTYKG-IVKSSBEG	158	
XYNC_STRLI	125	YGVVTRSEIIEYVYVDSWGTTEETGTYKG-IVKSSBEG	184	
PULPYME_L	153	ETLRVNCESIGK-IAETKQYVSVRRSTFSG--TISVSNKE	190	
XYNA_BACCI	137	ETLRVNCESIGK-IAETKQYVSVRRSTFSG--TISVSNKE	178	
XYNA_BACPU	151	ETLRVNCESIGK-IAETKQYVSVRRSTFSG--TISVSNKE	188	
XYNA_BACST	136	ETLRVNCESIGK-IAETKQYVSVRRSTFSG--TISVSNKE	176	
XYNA_BACSU	137	ETLRVNCESIGK-IAETKQYVSVRRSTFSG--TISVSNKE	178	
XYNA_CLOAB	184	ETLRVNCESIGK-IAETKQYVSVRRSTFSG--TISVSNKE	221	
XYNA_CLOSR	157	ETLRVNCESIGK-IAETKQYVSVRRSTFSG--TISVSNKE	194	
XYNB_STRLI	159	ETLRVNCESIGK-IAETKQYVSVRRSTFSG--TISVSNKE	196	
XYNC_STRLI	165	ETLRVNCESIGK-IAETKQYVSVRRSTFSG--TISVSNKE	204	
PULPYME_L	191	RAWENLGNMG-KMYEVALTVEGYSSGSAVYYSNTLRINGN	231	
XYNA_BACCI	179	RAWENLGNMG-KMYEVALTVEGYSSGSAVYYSNTLRINGN	213	
XYNA_BACPU	189	RAWENLGNMG-KMYEVALTVEGYSSGSAVYYSNTLRINGN	228	
XYNA_BACST	177	RAWENLGNMG-KMYEVALTVEGYSSGSAVYYSNTLRINGN	211	
XYNA_BACSU	179	RAWENLGNMG-KMYEVALTVEGYSSGSAVYYSNTLRINGN	213	
XYNA_CLOAB	222	AWAESKGLPLG-KMHETAFNIEGYSSGKADYNSMSINIGK	281	
XYNA_CLOSR	195	KOWERMGMRMG-KMYEVALTVEGYSSGSAVYYSNTLRINGN	235	
XYNB_STRLI	197	DAWARAGMPLGNFSYYMIMATIEGYSSGSAVYYSNTLRINGN	238	
XYNC_STRLI	205	DAWARAGMPLGNFSYYMIMATIEGYSSGSAVYYSNTLRINGN	240	

Figure 2

SUBSTITUTE SHEET (RULE 26)

PULPNS8-11	1	M R Q K K L T F I L A F L V C F A L T L P A E L L Q A Q I V T D N	33
PULPZYME_L	1	M R Q K K L T F I L A F L V C F A L T L P A E L L Q A Q I V T D N	33
PULPNS8-11	34	S I G N H D G Y D Y E F W K D S G G S G T M L E N H G G T F S A Q	66
PULPZYME_L	34	S I G N H D G Y D Y E F W K D S G G S G T M L E N H G G T F S A Q	66
PULPNS8-11	67	W N N V N N I L F R K G K K F N E T Q T H Q Q V G N M S I N Y G A	99
PULPZYME_L	67	W N N V N N I L F R K G K K F N E T Q T H Q Q V G N M S I N Y G A	99
PULPNS8-11	100	N E Q P N G N A Y L C V Y G W T V D P L V E Y Y E V D S W G N W R	132
PULPZYME_L	100	N E Q P N G N A Y L C V Y G W T V D P L V E Y Y E V D S W G N W R	132
PULPNS8-11	133	P P G A L P K G T I E V D G G E Y D E Y K H Q Q V N Q P S I Q G T	165
PULPZYME_L	133	P P G A L P K G T I E V D G G E Y D E Y E T L R V N Q P S I K G I	165
PULPNS8-11	166	A T E N Q Y W S I R Q S K R T S G T V T T A N F F E N A W A A E G M	198
PULPZYME_L	166	A T E K Q Y W S V R R S K R T S G T I S V S N F E R A W E N L G M	198
PULPNS8-11	199	N M G A F N Y Q I E V T E G Y Q S T G S A N V Y S N T L R E N G N	231
PULPZYME_L	199	N M G K M Y E V A L T V E G Y Q S S G S A N V Y S N T L R E N G N	231
PULPNS8-11	232	P L S T I S N D K S I T L D K N N	248
PULPZYME_L	232	P L S T I S N D K S I T L D K N N	248

Figure 3

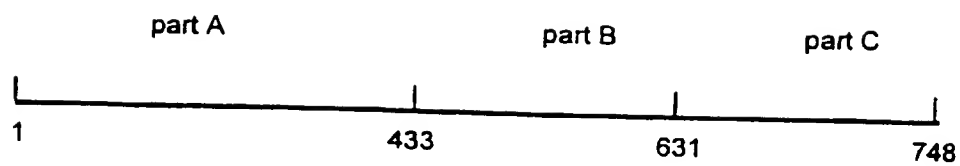


Figure 4